Atty. Docket No. MS303183.2

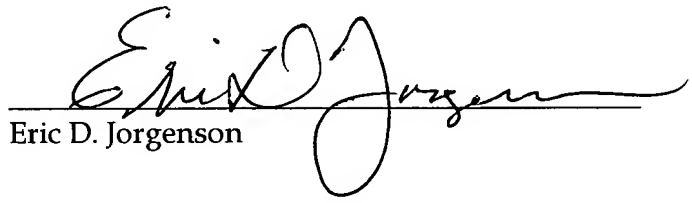
ARCHITECTURE FOR CONTROLLING A COMPUTER USING HAND GESTURES

by

Andrew D. Wilson and Nuria M. Oliver

MAIL CERTIFICATION

I hereby certify that the attached patent application (along with any other paper referred to as being attached or enclosed) is being deposited with the United States Postal Service on this date **December 1, 2003**, in an envelope as "Express Mail Post Office to Addressee" Mailing Label Number **EV330022246US** addressed to: Mail Stop: Patent Applications, Commissioner for Patents, P.O. Box 1450, Alexandria, Virginia 22313-1450



Eric D. Jorgenson

Title: ARCHITECTURE FOR CONTROLLING A COMPUTER USING HAND GESTURES

CROSS-REFERENCE TO RELATED APPLICATIONS

This application is a Continuation-in-Part of pending U.S. Patent Application Ser. No. 10/396,653 entitled "ARCHITECTURE FOR CONTROLLING A COMPUTER USING HAND GESTURES" filed March 25, 2003, the entirety of which is incorporated by reference.

TECHNICAL FIELD

The present invention relates generally to controlling a computer system, and more particularly to a system and method to implement alternative modalities for controlling computer programs and devices, and manipulating on-screen objects through the use of one or more body gestures, or a combination of gestures and supplementary signals.

BACKGROUND OF THE INVENTION

A user interface facilitates the interaction between a computer and computer user by enhancing the user's ability to utilize application programs. The traditional interface between a human user and a typical personal computer is implemented with graphical displays and is generally referred to as a graphical user interface (GUI). Input to the computer or particular application program is accomplished through the presentation of graphical information on the computer screen and through the use of a keyboard and/or mouse, trackball or other similar implements. Many systems employed for use in public areas utilize touch screen implementations whereby the user touches a designated area of a screen to effect the desired input. Airport electronic ticket check-in kiosks and rental car direction systems are examples of such systems. There are, however, many applications where the traditional user interface is less practical or efficient.

The traditional computer interface is not ideal for a number of applications. Providing stand-up presentations or other type of visual presentations to large audiences,

is but one example. In this example, a presenter generally stands in front of the audience and provides a verbal dialog in conjunction with the visual presentation that is projected on a large display or screen. Manipulation of the presentation by the presenter is generally controlled through use of awkward remote controls, which frequently suffer from inconsistent and less precise operation, or require the cooperation of another individual. Traditional user interfaces require the user either to provide input via the keyboard or to exhibit a degree of skill and precision more difficult to implement with a remote control than a traditional mouse and keyboard. Other examples include control of video, audio, and display components of a media room. Switching between sources, advancing fast fast-forward, rewinding, changing chapters, changing volume, *etc.*, can be very cumbersome in a professional studio as well as in the home. Similarly, traditional interfaces are not well suited for smaller, specialized electronic gadgets.

Additionally, people with motion impairment conditions find it very challenging to cope with traditional user interfaces and computer access systems. Such conditions include Cerebral Palsy, Muscular Dystrophy, Friedrich's Ataxia, and spinal injuries or disorders. These conditions and disorders are often accompanied by tremors, spasms, loss of coordination, restricted range of movement, reduced muscle strength, and other motion impairing symptoms.

Similar symptoms exist in the growing elderly segment of the population. As people age, their motor skills decline and impact the ability to perform many tasks. It is known that as people age, their cognitive, perceptual and motor skills decline, with negative effects in their ability to perform many tasks. The requirement to position a cursor, particularly with smaller graphical presentations, can often be a significant barrier for elderly or afflicted computer users. Computers can play an increasingly important role in helping older adults function well in society.

Graphical interfaces contribute to the ease of use of computers. WIMP (Window, Icon, Menu, Pointing device (or Pull-down menu)) interfaces allow fairly non-trivial operations to be performed with a few mouse motions and clicks. However, at the same time, this shift in the user interaction from a primarily text-oriented experience to a point-and-click experience has erected new barriers between people with disabilities and the computer. For example, for older adults, there is evidence that using the mouse can

be quite challenging. There is extensive literature demonstrating that the ability to make small movements decreases with age. This decreased ability can have a major effect on the ability of older adults to use a pointing device on a computer. It has been shown that even experienced older computer users move a cursor much more slowly and less accurately than their younger counterparts. In addition, older adults seem to have increased difficulty (as compared to younger users) when targets become smaller. For older computer users, positioning a cursor can be a severe limitation.

One solution to the problem of decreased ability to position the cursor with a mouse is to simply increase the size of the targets in computer displays, which can often be counter-productive since less information is being displayed, requiring more navigation. Another approach is to constrain the movement of the mouse to follow on-screen objects, as with sticky icons or solid borders that do not allow cursors to overshoot the target. There is evidence that performance with area cursors (possibly translucent) is better than performance with regular cursors for some target acquisition tasks.

One method to facilitate computer access for users with motion impairment conditions and for applications, in which the traditional user interfaces are cumbersome, is through use of perceptual user interfaces. Perceptual user interfaces utilize alternate sensing modalities, such as the capability of sensing physical gestures of the user, to replace or complement traditional input devices such as the mouse and keyboard. Perceptual user interfaces promise modes of fluid computer-human interaction that complement and/or replace the mouse and keyboard, particularly in non-desktop applications such as control for a media room.

One study indicates that adding a simple gesture-based navigation facility to web browsers can significantly reduce the time taken to carry out one of the most common actions in computer use, *i.e.*, using the "back" button (or function) to return to previously visited pages. Subjective ratings by users in experiments showed a strong preference for a "flick" system, where the users would flick the mouse left or right to go back or forward in the web browser.

In the simplest view, gestures play a symbolic communication role similar to speech, suggesting that for simple tasks gestures can enhance or replace speech

recognition. Small gestures near the keyboard or mouse do not induce fatigue as quickly as sustained whole arm postures. Previous studies indicate that users find gesture-based systems highly desirable, but that users are also dissatisfied with the recognition accuracy of gesture recognizers. Furthermore, experimental results indicate that a user's difficulty with gestures is in part due to a lack of understanding of how gesture recognition works. The studies highlight the ability of users to learn and remember gestures as an important design consideration.

Even when a mouse and keyboard are available, users may find it attractive to manipulate often-used applications while away from the keyboard, in what can be called a "casual interface" or "lean-back" posture. Browsing e-mail over morning coffee might be accomplished by mapping simple gestures to "next message" and "delete message".

Gestures can compensate for the limitations of the mouse when the display is several times larger than a typical display. In such a scenario, gestures can provide mechanisms to restore the ability to quickly reach any part of the display, where once a mouse was adequate with a small display. Similarly, in a multiple display scenario it is desirable to have a fast comfortable way to indicate a particular display. For example, the foreground object can be "bumped" to another display by gesturing in the direction of the target display.

However, examples of perceptual user interfaces to date are dependent on significant limiting assumptions. One type of perceptual user interface utilizes color models that make certain assumptions about the color of an object. Proper operation of the system is dependent on proper lighting conditions and can be negatively impacted when the system is moved from one location to another as a result of changes in lighting conditions, or simply when the lighting conditions change in the room. Factors that impact performance include sun light versus artificial light, florescent light versus incandescent light, direct illumination versus indirect illumination, and the like. Additionally, most attempts to develop perceptual user interfaces require the user to wear specialized devices such as gloves, headsets, or close-talk microphones. The use of such devices is generally found to be distracting and intrusive for the user.

Thus perceptual user interfaces have been slow to emerge. The reasons include heavy computational burdens, unreasonable calibration demands, required use of

intrusive and distracting devices, and a general lack of robustness outside of specific laboratory conditions. For these and similar reasons, there has been little advancement in systems and methods for exploiting perceptual user interfaces. However, as the trend towards smaller, specialized electronic gadgets continues to grow, so does the need for alternate methods for interaction between the user and the electronic device. Many of these specialized devices are too small and the applications unsophisticated to utilize the traditional input keyboard and mouse devices. Examples of such devices include TabletPCs, Media center PCs, kiosks, hand held computers, home appliances, video games, and wall sized displays, along with many others. In these, and other applications, the perceptual user interface provides a significant advancement in computer control over traditional computer interaction modalities.

In light of these findings, what is needed is to standardize a small set of easily learned gestures, the semantics of which are determined by application context. A small set of very simple gestures can offer significant bits of functionality where they are needed most. For example, dismissing a notification window can be accomplished by a quick gesture to the one side or the other, as in shooing a fly. Another example is gestures for "next" and "back" functionality found in web browsers, presentation programs (*e.g.*, PowerPoint™) and other applications. Note that in many cases the surface forms of these various gestures can remain the same throughout these examples, while the semantics of the gestures depends on the application at hand. Providing a small set of standard gestures eases problems users have in recalling how gestures are performed, and also allows for simpler and more robust signal processing and recognition processes.

SUMMARY OF THE INVENTION

The following presents a simplified summary of the invention in order to provide a basic understanding of some aspects of the invention. This summary is not an extensive overview of the invention. It is intended to neither identify key or critical elements of the invention nor delineate the scope of the invention. Its sole purpose is to present some concepts of the invention in a simplified form as a prelude to the more detailed description that is presented later.

The present invention disclosed and claimed herein, in one aspect thereof, comprises a system for controlling a computer using gestures. The system includes a 3-D imaging system that performs gesture recognition and interpretation based on a previous mapping of a plurality of hand poses and orientations to user commands for a given user. When the user is identified to the system, the imaging system images gestures presented by the user, performs a lookup for the user command associated with the captured image(s), and executes the user command(s) to effect control of the computer, programs, and connected devices.

In another aspect of the present invention, the system includes a wireless device worn by the person. The wireless device includes one or more sensors that measure at least velocity, acceleration, and orientation of the device. The corresponding signals are transmitted to a computer system, processed, and interpreted to determine an object at which the device is pointed and the action to be taken on the object. Once the signals have been interpreted, the computer is controlled to interact with the object, which object can be a device and/or system connected to the computer, and software running on the computer. In one application, the wireless device is used in a medical environment and worn on the head of a medical person allowing free use of the hands. Head movements facilitate control of the computer. In another multimodal approach, the person can also wear a wireless microphone to communicate voice signals to the computer separately or in combination with head movements for control thereof.

In yet another aspect of the present invention, a multimodal approach can be employed such that a person uses the wireless device in combination with the imaging capabilities of the 3-D imaging system.

In still another aspect of the present invention, the multimodal approach includes any combination of the 3-D imaging system, the wireless device, and vocalization to control the computer system and, hardware and software associated therewith. This approach finds application in a medical environment such as an operating room, for example.

In another aspect of the present invention, an engagement volume is employed in a medical environment such that one or both hands of the medical person are free to engage the volume and control the computer system, during, for example, a patient

operation. The volume is defined in space over the part of the patient undergoing the operation, and the hands of the medical person are used in the form of gestures to control the system for the presentation of medical information.

In accordance with another aspect thereof, the present invention facilitates adapting the system to the particular preferences of an individual user. The system and method allow the user to tailor the system to recognize specific hand gestures and verbal commands and to associate these hand gestures and verbal commands with particular actions to be taken. This capability allows different users, which may prefer to make different motions for a given command, the ability to tailor the system in a way most efficient for their personal use. Similarly, different users can choose to use different verbal commands to perform the same function.

In still another aspect of the present invention, the system employs a learning capability such that nuances of a user can be learned by the system and adapted to the user profile of gestures, vocalizations, etc.

The following description and the annexed drawings set forth in detail certain illustrative aspects of the invention. These aspects are indicative, however, of but a few of the various ways in which the principles of the invention can be employed and the present invention is intended to include all such aspects and their equivalents. Other advantages and novel features of the invention will become apparent from the following detailed description of the invention when considered in conjunction with the drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 illustrates a system block diagram of components of the present invention for controlling a computer and/or other hardware/software peripherals interfaced thereto.

FIG. 2 illustrates a schematic block diagram of a perceptual user interface system, in accordance with an aspect of the present invention.

FIG. 3 illustrates a flow diagram of a methodology for implementing a perceptual user interface system, in accordance with an aspect of the present invention.

FIG. 4 illustrates a flow diagram of a methodology for determining the presence of moving objects within images, in accordance with an aspect of the present invention.

FIG. 5 illustrates a flow diagram of a methodology for tracking a moving object within an image, in accordance with an aspect of the present invention.

FIG. 6 illustrates a disparity between two video images captured by two video cameras mounted substantially parallel to each other for the purpose of determining the depth of objects, in accordance with an aspect of the present invention.

FIG. 7 illustrates an example of the hand gestures that the system can recognize and the visual feedback provided through the display, in accordance with an aspect of the present invention.

FIG. 8 illustrates an alternative embodiment wherein a unique icon is displayed in association with a name of a specific recognized command, in accordance with an aspect of the present invention.

FIGs. 9A and 9B illustrate an engagement plane and volume of both single and multiple monitor implementations, in accordance with an aspect of the present invention.

FIG. 10 illustrates a briefing room environment where gestures are utilized to control a screen projector via a computer system configured in accordance with an aspect of the present invention.

FIG. 11 illustrates a block diagram of a computer system operable to execute the present invention.

FIG. 12 illustrates a network implementation of the present invention.

FIG. 13 illustrates a medical operating room system that uses the engagement volume in accordance with the present invention.

FIG. 14 illustrates a medical operating room environment in which a computer control system with wireless control device is employed in accordance with the present invention.

FIG. 15 illustrates a flowchart of a process from the perspective of the person for using the system of FIG. 14.

FIG. 16 illustrates a flowchart of a process from the perspective of the system of FIG. 14.

FIG. 17 illustrates a medical environment in which a 3-D imaging computer control system is employed to process hand (or body) gestures in accordance with the present invention.

FIG. 18 illustrates a flowchart of a process from the perspective of the person for using the system of FIG. 17.

FIG. 19 illustrates a flowchart of a process from the perspective of the system of FIG. 17.

FIG. 20 illustrates a medical environment in which a 3-D imaging computer control system is employed with the remote control device to process hand (or body) gestures and control the system in accordance with the present invention.

FIG. 21A illustrates a sample one-handed and two-handed gestures that can be used to control the operation computing system in accordance with the present invention.

FIG. 21B illustrates an additional sample one-handed gestures and sequenced one-handed gestures that can be used to control the operation computing system in accordance with the present invention.

FIG. 21C illustrates additional sample one-handed gestures that can be used to control the operation computing system in accordance with the present invention.

FIG. 21D illustrates additional sample one-handed gestures used in combination with voice commands that can be used to control the operation computing system in accordance with the present invention.

FIG. 21E illustrates additional sample one-handed gestures used in combination with voice commands and gaze signals that can be used to control the operation computing system in accordance with the present invention.

DETAILED DESCRIPTION OF THE INVENTION

As used in this application, the terms "component" and "system" are intended to refer to a computer-related entity, either hardware, a combination of hardware and software, software, or software in execution. For example, a component may be, but is not limited to being, a process running on a processor, a processor, an object, an executable, a thread of execution, a program, and/or a computer. By way of illustration, both an application running on a server and the server can be a component. One or more components may reside within a process and/or thread of execution and a component may be localized on one computer and/or distributed between two or more computers.

The present invention relates to a system and methodology for implementing a perceptual user interface comprising alternative modalities for controlling computer programs and manipulating on-screen objects through hand gestures or a combination of hand gestures and/or verbal commands. A perceptual user interface system is provided that tracks hand movements and provides for the control of computer programs and manipulation of on-screen objects in response to hand gestures performed by the user. Similarly, the system provides for the control of computer programs and manipulation of on-screen objects in response to verbal commands spoken by the user. Further, the gestures and/or verbal commands can be tailored by a particular user to suit that user's personal preferences. The system operates in real time and is robust, light-weight and responsive. The system provides a relatively inexpensive capability for the recognition of hand gestures and verbal commands.

Referring now to FIG. 1, there is illustrated a system block diagram of components of the present invention for controlling a computer and/or other hardware/software peripherals interfaced thereto. The system 100 includes a tracking component 102 for detecting and tracking one or more objects 104 through image capture utilizing cameras (not shown) or other suitable conventional image-capture devices. The cameras operate to capture images of the object(s) 104 in a scene within the image capture capabilities of the cameras so that the images can be further processed to not only detect the presence of the object(s) 104, but also to detect and track object(s) movements. It is appreciated that in more robust implementations, object characteristics such as object features and object orientation can also be detected, tracked, and processed. The object(s) 104 of the present invention include basic hand movements created by one or more hands of a system user and/or other person selected for use with the disclosed system. However, in more robust system implementations, such objects can include many different types of objects with object characteristics, including hand gestures each of which have gesture characteristics including but not limited to, hand movement, finger count, finger orientation, hand rotation, hand orientation, and hand pose (*e.g.*, opened, closed, and partially closed).

The tracking component 102 interfaces to a control component 106 of the system 100 that controls all onboard component processes. The control component 106

interfaces to a seeding component 108 that seeds object hypotheses to the tracking component based upon the object characteristics.

The object(s) 104 are detected and tracked in the scene such that object characteristic data is processed according to predetermined criteria to associate the object characteristic data with commands for interacting with a user interface component 110. The user interface component 110 interfaces to the control component 106 to receive control instructions that affect presentation of text, graphics, and other output (*e.g.*, audio) provided to the user via the interface component 110. The control instructions are communicated to the user interface component 110 in response to the object characteristic data processed from detection and tracking of the object(s) within a predefined engagement volume space 112 of the scene.

A filtering component 114 interfaces to the control component 106 to receive filtering criteria in accordance with user filter configuration data, and to process the filtering criteria such that tracked object(s) of respective object hypotheses are selectively removed from the object hypotheses and/or at least one hypothesis from a set of hypotheses within the volume space 112 and the scene. Objects are detected and tracked either within the volume space 112 or outside the volume space 112. Those objects outside of the volume space 112 are detected, tracked, and ignored, until entering the volume space 112.

The system 100 also receives user input *via* input port(s) 116 such as input from pointing devices, keyboards, interactive input mechanisms such as touch screens, and audio input devices.

The subject invention (*e.g.*, in connection with object detection, tracking, and filtering) can employ various artificial intelligence based schemes for carrying out various aspects of the subject invention. For example, a process for determining which object is to be selected for tracking can be facilitated *via* an automatic classification system and process. Such classification can employ a probabilistic and/or statistical-based analysis (*e.g.*, factoring into the analysis utilities and costs) to prognose or infer an action that a user desires to be automatically performed. For example, a support vector machine (SVM) classifier can be employed. Other classification approaches include Bayesian networks, decision trees, and probabilistic classification models providing

different patterns of independence can be employed. Classification as used herein also is inclusive of statistical regression that is utilized to develop models of priority.

As will be readily appreciated from the subject specification, the subject invention can employ classifiers that are explicitly trained (*e.g.*, *via* a generic training data) as well as implicitly trained (*e.g.*, *via* observing user behavior, receiving extrinsic information) so that the classifier(s) is used to automatically determine according to a predetermined criteria which object(s) should be selected for tracking and which objects that were being tracked are now removed from tracking. The criteria can include, but is not limited to, object characteristics such as object size, object speed, direction of movement, distance from one or both cameras, object orientation, object features, and object rotation. For example, with respect to SVM's which are well understood - it is to be appreciated that other classifier models can also be utilized such as Naive Bayes, Bayes Net, decision tree and other learning models - SVM's are configured *via* a learning or training phase within a classifier constructor and feature selection module. A classifier is a function that maps an input attribute vector, $\mathbf{x} = (x1, x2, x3, x4,\ldots, xn)$, to a confidence that the input belongs to a class - that is, $f(\mathbf{x}) = confidence(class)$. In the case of object identification and tracking, for example, attributes include various sizes of the object, various orientations of the object, object colors, and the classes are categories or areas of interest (*e.g.*, object type, and object pose).

Referring now to FIG. 2, there is illustrated a schematic block diagram of a perceptual user interface system, in accordance with an aspect of the present invention. The system comprises a computer 200 with a traditional keyboard 202, input pointing device (*e.g.*, a mouse) 204, microphone 206, and display 208. The system further comprises at least one video camera 210, at least one user 212, and software 214. The exemplary system of FIG. 2 is comprised of two video cameras 210 mounted substantially parallel to each other (that is, the rasters are parallel) and the user 212. The first camera is used to detect depth of the object from the camera and track the object, and the second camera is used for determining at least the depth (or distance) of the object from the camera(s). The computer 200 is operably connected to the keyboard 202, mouse 204 and display 208. Video cameras 210 and microphone 206 are also operably connected to computer 200. The video cameras 210 "look" towards the user 212 and

may point downward to capture objects within the volume defined above the keyboard and in front of the user. User 212 is typically an individual that is capable of providing hand gestures, holding objects in a hand, verbal commands, and mouse and/or keyboard input. The hand gestures and/or object(s) appear in video images created by the video cameras 210 and are interpreted by the software 214 as commands to be executed by computer 200. Similarly, microphone 206 receives verbal commands provided by user 212, which are in turn, interpreted by software 214 and executed by computer 200. User 212 can control and operate various application programs on the computer 200 by providing a series of hand gestures or a combination of hand gestures, verbal commands, and mouse/keyboard input. The system can track any object presented in the scene in front or it. The depth information is used to "segment" the interacting object from the rest of the scene. The capability to exploit any sort of moving object in the scene is important at least with respect to accessibility scenarios.

In view of the foregoing structural and functional features described above, methodologies in accordance with various aspects of the present invention will be better appreciated with reference to FIGs. 3-5. While, for purposes of simplicity of explanation, the methodologies of FIGs. 3-5 are shown and described as executing serially, it is to be understood and appreciated that the present invention is not limited by the illustrated order, as some aspects could, in accordance with the present invention, occur in different orders and/or concurrently with other aspects from that shown and described herein. Moreover, not all illustrated features may be required to implement a methodology in accordance with an aspect the present invention.

Accordingly, FIG. 3 is a flow diagram that illustrates a high level methodology for detecting the user's hand, tracking movement of the hand and interpreting commands in accordance with an aspect of the invention. While, for purposes of simplicity of explanation, the methodologies shown here and below are described as a series of acts, it is to be understood and appreciated that the present invention is not limited by the order of acts, as some acts may, in accordance with the present invention, occur in different orders and/or concurrently with other acts from that shown and described herein. For example, those skilled in the art will understand and appreciate that a methodology could alternatively be represented as a series of interrelated states or events, such as in a state

diagram. Moreover, not all illustrated acts may be required to implement a methodology in accordance with the present invention.

The methodology begins at 300 where video images are scanned to determine whether any moving objects exist within the field of view (or scene) of the cameras. The system is capable of running one or more object hypothesis models to detect and track objects, whether moving or not moving. In one embodiment, the system runs up to and including six object hypotheses. If more than one object is detected as a result of the multiple hypotheses, the system drops one of the objects if the distance from any other object falls below a threshold distance, for example, five inches. It is assumed that the two hypotheses are redundantly tracking the same object, and one of the hypotheses is removed from consideration. At 302, if NO, no moving object(s) have been detected, and flow returns to 300 where the system continues to scan the current image for moving objects. Alternatively, if YES, object movement has been detected, and flow continues from 302 to 304 where it is determined whether or not one or more moving objects are within the engagement volume. It is appreciated that the depth of the object may be determined before determination of whether the object is within the engagement volume.

The engagement volume is defined as a volume of space in front of the video cameras and above the keyboard wherein the user is required to introduce the hand gestures (or object(s)) in order to utilize the system. A purpose of the engagement volume is to provide a means for ignoring all objects and/or gestures in motion except for those intended by the user to effect control of the computer. If a moving object is detected at 302, but is determined not to be within the engagement volume, then the system dismisses the moving object as not being a desired object to track for providing commands. Flow then loops back to the input of 300 to scan for more objects. However, if the moving object is determined to be within the engagement volume, then the methodology proceeds to 306. However, new objects are seeded only when it is determined that the new object is a sufficient distance away from an existing object that is being tracked (in 3-D). At 306, the system determines the distance of each moving object from the video cameras. At 308, if more than one moving object is detected within the engagement volume, then the object closest to the video camera(s) is selected as the desired command object. If by the given application context the user is predisposed to

use hand gestures towards the display, the nearest object hypotheses will apply to the hands. In other scenarios, more elaborate criteria for object selection may be used. For example, an application may select a particular object based upon its quality of movement over time. Additionally, a two-handed interaction application may select an object to the left of the dominant hand (for right handed users) as the non-dominant hand. The command object is the object that has been selected for tracking, the movements of which will be analyzed and interpreted for gesture commands. The command object is generally the user's dominant hand. Once the command object is selected, its movement is tracked, as indicated at 310.

At 312, the system determines whether the command object is still within the engagement volume. If NO, the object has moved outside the engagement volume, and the system dismisses the object hypothesis and returns to 300 where the current image is processed for moving objects. If NO, the object is still within the engagement volume, and flow proceeds to 314. At 314, the system determines whether the object is still moving. If no movement is detected, flow is along the NO path returning to 300 to process the current camera images for moving objects. If however, movement is detected, then flow proceeds from 314 to 316. At 316, the system analyzes the movements of the command object to interpret the gestures for specific commands. At 318, it is determined whether the interpreted gesture is a recognized command. If NO, the movement is not interpreted as a recognized command, and flow returns to 310 to continue tracking the object. However, if the object movement is interpreted as a recognized command, flow is to 320 where the system executes the corresponding command. After execution thereof, flow returns to 310 to continue tracking the object. This process may continually execute to detect and interpret gestures.

In accordance with an aspect of the invention, algorithms used to interpret gestures are kept to simple algorithms and are performed on sparse ("lightweight") images to limit the computational overhead required to properly interpret and execute desired commands in real time. In accordance with another aspect of the invention, the system is able to exploit the presence of motion and depth to minimize computational requirements involved in determining objects that provide gesture commands.

Referring now to FIG. 4, there is illustrated a flow diagram of a methodology for determining the presence of moving objects within video images created by one or more video sources, in accordance with an aspect of the present invention. The methodology exploits the notion that attention is often drawn to objects that move. At 400, video data is acquired from one or more video sources. Successive video images are selected from the same video source, and motion is detected by comparing a patch of a current video image, centered on a given location, to a patch from the previous video image centered on the same location. At 402, a video patch centered about a point located at (u_1,v_1), and (u_2,v_2) is selected from successive video images I_1 and I_2, respectively. A simple comparison function is utilized wherein the sum of the absolute differences (SAD) over square patches in two images is obtained. For a patch from image I_1 centered on pixel location (u_1,v_1) and a patch in image I_2 centered on (u_2,v_2), the image comparison function is defined as $SAD(I_1,u_1,v_1,I_2,u_2,v_2)$ as:

$$\sum_{-\frac{D}{2}\le i,j\le\frac{D}{2}} \left| I_1(u_1+i,v_1+j) - I_2(u_2+i,v_2+j) \right|$$

where $I(u,v)$ refers to the pixel at (u,v), D is the patch width, and the absolute difference between two pixels is the sum of the absolute differences taken over all available color channels. Regions in the image that have movement can be found by determining points (u,v) such that function $SAD(I_{t-1},u_{t-1},v_{t-1},I_t,u_t,v_t) > t$, where the subscript refers to the image at time t, and t is a threshold level for motion. At 404, a comparison is made between patches from image I_1 and I_2 using the sum of the absolute difference algorithm. At 406, the result of the sum of the absolute difference algorithm is compared to a threshold value to determine whether a threshold level of motion exists within the image patch. If $SAD = t$, no sufficient motion exists, and flow proceeds to 410. If at 406, $SAD > t$, then sufficient motion exists within the patch, and flow is to 408 where the object is designated for continued tracking. At 410, the system determines whether the current image patch is the last patch to be examined within the current image. If NO, the methodology returns to 402 where a new patch is selected. If YES, then the system returns to 400 to acquire a new video image from the video source.

To reduce the computational load, the SAD algorithm is computed on a sparse regular grid within the image. In one embodiment, the sparse regular grid is based on sixteen pixel centers. When the motion detection methodology determines that an object has sufficient motion, then the system tracks the motion of the object. Again, in order to limit (or reduce) the computational load, a position prediction algorithm is used to predict the next position of the moving object. In one embodiment, the prediction algorithm is a Kalman filter. However, it is to be appreciated that any position prediction algorithm can be used.

Note that the image operations may use the same SAD function on image patches, which allows for easy SIMD (Single-Instruction Multiple-Data, which architectures are essential in the parallel world of computers) optimization of the algorithm's implementation, which in turn allows it to run with sufficiently many trackers while still leaving CPU time to the user.

The process of seeding process hypotheses based upon motion may place more than one hypothesis on a given moving object. One advantage of this multiple hypothesis approach is that a simple, fast, and imperfect tracking algorithm may be used. Thus if one tracker fails, another may be following the object of interest. Once a given tracker has been seeded, the algorithm updates the position of the object being followed using the same function over successive frames.

Referring now to FIG. 5, there is illustrated a flow diagram of a methodology for tracking a moving object within an image, in accordance with an aspect of the present invention. The methodology begins at 500 where, after the motion detection methodology has identified the location of a moving object to be tracked, the next position of the object is predicted. Once identified, the methodology utilizes a prediction algorithm to predict the position of the object in successive frames. The prediction algorithm limits the computational burden on the system. In the successive frames, the moving object should be at the predicted location, or within a narrow range centered on the predicted location. At 502, the methodology selects a small pixel window (*e.g.*, ten pixels) centered on the predicted location. Within this small window, an algorithm executes to determine the actual location of the moving object. At 504, the new position is determined by examining the sum of the absolute difference algorithm over successive

video frames acquired at time t and time $t-1$. The actual location is determined by finding the location (u_t, v_t) that minimizes:

$$SAD(I_{t-1},u_{t-1},v_{t-1},I_t,u_t,v_t),$$

where I_t refers to the image at time t, I_{t-1} refers to the image at time $t-1$, and where (u_t, v_t) refers to the location at time t. Once determined, the actual position is updated, at 506. At 508, motion characteristics are evaluated to determine whether the motion is still greater that the threshold level required. What is evaluated is not only the SAD image-based computation, but also movement of the object over time. The movement parameter is the average movement over a window of time. Thus if the user pauses the object or hand for a short duration of time, it may not be dropped from consideration. However, if the duration of time for the pause is still longer such that it exceeds a predetermined average time parameter, the object will be dropped. If YES, the motion is sufficient, and flow returns to 500 where a new prediction for the next position is determined. If NO, the object motion is insufficient, and the given object is dropped from being tracked, as indicated by flow to 510. At 512, flow is to 430 of FIG. 4 to select a new patch in the image from which to analyze motion.

When determining the depth information of an object (*i.e.*, the distance from the object to the display or any other chosen reference point), a lightweight sparse stereo approach is utilized in accordance with an aspect of the invention. The sparse stereo approach is a region-based approach utilized to find the disparity at only locations in the image corresponding to the object hypothesis. Note that in the stereo matching process, it is assumed that both cameras are parallel (in rasters). Object hypotheses are supported by frame-to-frame tracking through time in one view and stereo matching across both views. A second calibration issue is the distance between the two cameras (*i.e.*, the baseline), which must be considered to recover depth in real world coordinates. In practice, both calibration issues maybe dealt with automatically by fixing the cameras on a prefabricated mounting bracket or semi-automatically by the user presenting objects at a known depth in a calibration routine that requires a short period of time to complete. The

accuracy of the transform to world coordinates is improved by accounting for lens distortion effects with a static, pre-computed calibration procedure for a given camera.

Binocular disparity is the primary means for recovering depth information from two or more images taken from different viewpoints. Given the two-dimensional position of an object in two views, it is possible to compute the depth of the object. Given that the two cameras are mounted parallel to each other in the same horizontal plane, and given that the two cameras have a focal length *f*, the three-dimensional position *(x,y,z)* of an object is computed from the positions of the object in both images (u_l, v_l) and (u_r, v_r) by the following perspective projection equations:

$$u = u_r = f\frac{x}{z};$$

$$v = v_r = f\frac{y}{z};$$

$$d = u_r - u_l = f\frac{b}{z};$$

where the disparity, *d*, is the shift in location of the object in one view with respect to the other, and is related to the baseline *b*, the distance between the two cameras.

The vision algorithm performs 3-dimensional (3-D) tracking and 3-D depth computations. In this process, each object hypothesis is supported only by consistency of the object movement in 3-D. Unlike many conventional computer vision algorithms, the present invention does not rely on fragile appearance models such as skin color models or hand image templates, which are likely invalidated when environmental conditions change or the system is confronted with a different user.

Referring now to FIG. 6, there is illustrated a disparity between two video images captured by two video cameras mounted substantially parallel to each other for the purpose of determining the depth of objects, in accordance with an aspect of the present invention. In FIG. 6, a first camera 600 and a second camera 602 (similar to cameras 210) are mounted substantially parallel to each other in the same horizontal plane and laterally aligned. The two cameras (600 and 602) are separated by a distance 604 defined between the longitudinal focal axis of each camera lens, also known as the baseline, *b*. A

first video image 606 is the video image from the first camera 600 and a second video image 608 is the video image from the second camera 602. The disparity *d* (also item number 610), or shift in the two video images (606 and 608), can be seen by looking to an object 612 in the center of the first image 606, and comparing the location of that object 612 in the first image 606 to the location of that same object 612 in the second image 608. The disparity 610 is illustrated as the difference between a first vertical centerline 614 of the first image 606 that intersects the center of the object 612, and a second vertical centerline 616 of the second image 608. In the first image 606, the object 612 is centered about the vertical centerline 614 with the top of the object 612 located at point *(u,v)*. In the second image 608, the same point *(u,v)* of the object 612 is located at point *(u-d,v)* in the second image 608, where d is the disparity 610, or shift in the object from the first image 606 with respect to the second image 610. Given disparity *d*, a depth *z* can be determined. As will be discussed, in accordance with one aspect of the invention, the depth component *z* is used in part to determine if an object is within the engagement volume, where the engagement volume is the volume within which objects will be selected by the system.

In accordance with another aspect of the present invention, a sparse stereo approach is utilized in order to limit computational requirements. The sparse stereo approach is that which determines disparity *d* only at the locations in the image that corresponds to a moving object. For a given point *(u,v)* in the image, the value of disparity *d* is found such that the sum of the absolute differences over a patch in the first image 606 (*i.e.*, a left image I_L) centered on *(u,v)* and a corresponding patch in the second image 608 (*i.e.*, a right image I_R) centered on *(u-d,v)*, is minimized, *i.e.*, the disparity value *d* that minimizes $SAD(I_l,u-d,v,I_r,u,v)$. If an estimate of depth *z* is available from a previous time, then in order to limit computational requirements, the search for the minimal disparity *d* is limited to a range consistent with the last known depth *z*.

In accordance with another aspect of the invention, the search range may be further narrowed by use of an algorithm to predict the objects new location. In one embodiment, the prediction is accomplished by utilization of a Kalman filter.

The depth *z* can also be computed using traditional triangulation techniques. The sparse stereo technique is used when the system operation involves detecting moving

objects within a narrow range in front of the display, *e.g.*, within twenty inches. In such cases, the two video cameras are mounted in parallel and can be separated by a distance equal to the approximate width of the display, or a even smaller distance that approximates a few inches. However, when the system is implemented in a larger configuration, the distance between the two video cameras may be much greater. In such cases, traditional triangulation algorithms are used to determine the depth.

The foregoing discussion has focused on some details of the methodologies associated with locating and tracking an object to effect execution of corresponding and specified commands. An overview follows as to how these capabilities are implemented in one exemplary system.

Referring now to FIG. 7, there is illustrated an example of gestures that the system recognizes, and further illustrates visual feedback provided to the system through the display. A user 700 gives commands by virtue of different hand gestures 702 and/or verbal commands 704. The gestures 702 are transmitted to a system computer (not shown) as part of the video images created by a pair of video cameras (706 and 708). Verbal and/or generally, audio commands, are input to the system computer through a microphone 710. Typical GUI windows 712, 714, and 716 are displayed in a layered presentation in an upper portion of display 718 while a lower portion of display 718 provides visual graphic feedback of in the form of icons 720, 722, 724, and 726 of some of the gestures 702 recognized by the system.

In one example, the hand icon 720 is displayed when a corresponding gesture 728 is recognized. The name of the recognized command (Move) is also then displayed below the icon 720 to provide additional textual feedback to the user 700. Move and Raise commands may be recognized by dwelling on the window for a period of time. There is also a "flick" or "bump" command to send a window from one monitor to another monitor, in a multiple monitor configuration. This is controlled by moving the hand (or object) to the left or right, and is described in greater detail hereinbelow with respect to FIG. 9B. There are at least two ways to effect a Move; by speech recognition when voicing the word "Move", or phrase "Move Window", or any other associated voice command(s); and, by using the dwelling technique. It is appreciated that where more robust image capture and imaging processing systems are implemented, the pose of

the hand may be mapped to any functionality, as described in greater detail below. Moreover, the shape of the hand icon may be changed in association with the captured hand pose to provide visual feedback to the user that the correct hand pose is being processed. However, as a basic implementation, the hand icon is positioned for selecting the window for interaction, or to move the window, or effect scrolling.

A Scroll command may be initiated first by voicing a corresponding command that is processed by speech recognition, and then using the hand (or object) to commence scrolling of the window by moving the hand (or object) up and down for the desired scroll direction.

In another example, the single displayed hand icon 720 is presented for all recognized hand gestures 702, however, the corresponding specific command name is displayed below the icon 720. Here, the same hand icon 720 is displayed in accordance with four different hand gestures utilized to indicate four different commands: Move, Close, Raise, and Scroll.

In still another aspect of the present invention, a different hand shaped icon is used for each specific command and the name of the command is optionally displayed below the command. In yet another embodiment, audio confirmation is provided by the computer, in addition to the displayed icon and optional command name displayed below the icon.

As previously mentioned, FIG. 7 illustrates the embodiment where a single hand shaped icon 720 is used, and the corresponding command recognized by the system is displayed below the icon 720. For example, when the system recognizes, either by virtue of gestures (with hand and/or object) and or verbal commands, the command to move a window, the icon 720 and corresponding command word "MOVE" are displayed by the display 718. Similarly, when the system recognizes a command to close a window, the icon 720 and corresponding command word "CLOSE" may be displayed by the display 718. Additional examples include, but are not limited to, displaying the icon 720 and corresponding command word "RAISE" when the system recognizes a hand gesture to bring a GUI window forward. When the system recognizes a hand gesture corresponding to a scroll command for scrolling a GUI window, the icon 720 and command word "SCROLL" are displayed by the display 718.

It is to be appreciated that the disclosed system may be configured to display any number and type of graphical icons in response to one or more hand gestures presented by the system user. Additionally, audio feedback may be used such that a beep or tone may be presented in addition to or *in lieu* of the graphical feedback. Furthermore the graphical icon may be used to provide feedback in the form of a color, combination of colors, and/or flashing color or colors. Feedback may also be provided by flashing a border of the selected window, the border in the direction of movement. For example, if the window is to be moved to the right, the right window border could be flashed to indicate the selected direction of window movement. In addition to or separate from, a corresponding tone frequency or any other associated sound may be emitted to indicate direction of movement, *e.g.*, an upward movement would have and associated high pitch and a downward movement would have a low pitch. Still further, rotational aspects may be provided such that movement to the left effects a counterclockwise rotation of a move icon, or perhaps a leftward tilt in the GUI window in the direction of movement.

Referring now to FIG. 8, there is illustrated an alternative embodiment wherein a unique icon is displayed in association with a name of a specific recognized command, in accordance with an aspect of the present invention. Here, each icon-word pair is unique for each recognized command. Icon-word pairs 800, 802, 804, and 806 for the respective commands "MOVE", "CLOSE", "RAISE", and "SCROLL", are examples of visual feedback capabilities that can be provided.

The system is capable of interpreting commands based on interpreting hand gestures, verbal commands, or both in combination. A hand is identified as a moving object by the motion detection algorithms and the hand movement is tracked and interpreted. In accordance with one aspect of the invention, hand gestures and verbal commands are used cooperatively. Speech recognition is performed using suitable voice recognition applications, for example, Microsoft SAPI 5.1, with a simple command and control grammar. However, it is understood that any similar speech recognition system can be used. An inexpensive microphone is placed near the display to receive audio input. However, the microphone can be placed at any location insofar as audio signals can be received thereinto and processed by the system.

Following is an example of functionality that is achieved by combining hand gesture and verbal modalities. Interaction with the system can be initiated by a user moving a hand across an engagement plane and into an engagement volume.

Referring now to FIG. 9A, there is illustrated the engagement plane and engagement volume for a single monitor system of the present invention. A user 900 is located generally in front of a display 902, which is also within the imaging capabilities of a pair of cameras (906 and 908). A microphone 904 (similar to microphones 206 and 710) is suitably located such that user voice signals are input for processing, *e.g.*, in front of the display 902. The cameras (906 and 908, similar to cameras 200 and, 706 and 708) are mounted substantially parallel to each other and on a horizontal plane above the display 902. The two video cameras (906 and 908) are separated by a distance that provides optimum detection and tracking for the given cameras and the engagement volume. However, it is to be appreciated that cameras suitable for wider fields of view, higher resolution, may be placed further apart on a plane different from the top of the display 902, for example, lower and along the sides of the display facing upwards, to capture gesture images for processing in accordance with novel aspects of the present invention. In accordance therewith, more robust image processing capabilities and hypothesis engines can be employed in the system to process greater amounts of data.

Between the display 902 and the user 900 is a volume 910 defined as the engagement volume. The system detects and tracks objects inside and outside of the volume 910 to determine the depth of one or more objects with respect to the engagement volume 910. However, those objects determined to be of a depth that is outside of the volume 910 will be ignored. As mentioned hereinabove, the engagement volume 910 is typically defined to be located where the hands and/or objects in the hands of the user 900 are most typically situated, *i.e.*, above a keyboard of the computer system and in front of the cameras (906 and 908) between the user 900 and the display 902 (provided the user 900 is seated in front of the display on which the cameras (906 and 908) are located). However, is it appreciated that the user 900 may be standing while controlling the computer, which requires that the volume 910 be located accordingly to facilitate interface interaction. Furthermore, the objects may include not only the hand(s) of the

user, or objects in the hand(s), but other parts of the body, such as head, torso movement, arms, or any other detectable objects. This is described in greater detail hereinbelow.

A plane 912 defines a face of the volume 910 that is closest to the user 900, and is called the engagement plane. The user 900 may effect control of the system by moving a hand (or object) through the engagement plane 912 and into the engagement volume 910. However, as noted above, the hand of the user 900 is detected and tracked even when outside the engagement volume 910. However, it would be ignored when outside of the engagement volume 910 insofar as control of the computer is concerned. When the object is moved across the engagement plane 912, feedback is provided to the user in the form of displaying an alpha-blended icon on the display (*e.g.*, an operating system desktop). The icon is designed to be perceived as distinct from other desktop icons and may be viewed as an area cursor. The engagement plane 912 is positioned such that the user's hands do not enter it during normal use of the keyboard and mouse. When the system engages the hand or object, the corresponding hand icon displayed on the desktop is moved to reflect the position of the tracked object (or hand).

The engagement and acquisition of the moving hand (or object) is implemented in the lightweight sparse stereo system by looking for the object with a depth that is less than a predetermined distance value. Any such object will be considered the command object until it is moved out of the engagement volume 910, for example, behind the engagement plane 912, or until the hand (or object) is otherwise removed from being a tracked object. In one example, the specified distance is twenty inches.

In operation, the user 900 moves a hand through the engagement plane 912 and into the engagement volume 910 established for the system. The system detects the hand, tracks the hand as the hand moves from outside of the volume 910 to the inside, and provides feedback by displaying a corresponding hand shaped icon on the display 902. The open microphone 904 placed near the display 902 provides means for the user 900 to invoke one or more verbal commands in order to act upon the selected window under the icon. The window directly underneath the hand shaped icon is the selected window. When a spoken and/or audio command is input to and understood by the system, the interpreted command is displayed along with the hand shaped icon. For example, in one embodiment, by speaking the word "Move", the user may initiate the continuous (or

stepped) movement of the window under the hand shaped icon to follow the movement of the user's hand. The user 900 causes the selected window to move up or down within the display 902 by moving the hand up or down. Lateral motion is also similarly achieved. Movement of the window is terminated when the user hand is moved across the engagement plane 912 and out of the engagement volume 910. Other methods of termination include stopping movement of the hand (or object) for an extended period of time, which is processed by the system as a command to drop the associated hypothesis. Furthermore, as described hereinabove, the Move command may be invoked by dwelling the hand on the window for a period of time, followed by hand motion to initiate the direction of window movement.

Alternatively, the user may speak the word "Release" and the system will stop moving the selected window in response to the user's hand motion. Release may also be accomplished by dwelling a bit longer in time while in Move, and/or Scroll modes. The user 900 may also act upon a selected window with other actions. By speaking the words, "Close", "Minimize", or "Maximize" the selected window is respectively closed, minimized or maximized. By speaking the word "Raise", the selected window is brought to the foreground, and by speaking "Send to Back", the selected window is sent behind (to the background) all other open windows. By speaking "Scroll", the user initiates a scrolling mode on the selected window. The user may control the rate of the scroll by the position of the hand. The hand shaped icon tracks the user's hand position, and the rate of the scrolling of the selected window is proportional to the distance between the current hand icon position and the position of the hand icon at the time the scrolling is initiated. Scrolling can be terminated by the user speaking "Release" or by the user moving their hand behind the engagement plane and out of the engagement volume. These are just a few examples of the voice recognition perceptual computer control capabilities of the disclosed architecture. It is to be appreciated that these voiced commands may also be programmed for execution in response to one or more object movements in accordance with the present invention.

In accordance with another aspect of the invention, dwell time can be used as a modality to control windows *in lieu* of, or in addition to, verbal commands and other disclosed modalities. Dwell time is defined as the time, after having engaged the system,

that the user holds their hand position stationary such that the system hand shaped icon remains over a particular window. For example, by dwelling on a selected window for a short period of time (*e.g.*, two seconds), the system can bring the window to the foreground of all other open windows (*i.e.*, a RAISE command). Similarly, by dwelling a short time longer (*e.g.*, four seconds), the system will grab (or select for dragging) the window, and the user causes the selected window to move up or down within the display by moving a hand up or down (*i.e.*, a MOVE command). Lateral motion is also similarly achieved. Additional control over GUI windows can be accomplished in a similar fashion by controlling the dwell time of the hand shaped icon over the open window.

In accordance with a more robust aspect of the invention, hand gestures are interpreted by hand motion or by pattern recognition. For example, the user can bring the window to the front (or foreground), on top of all other open windows by moving a hand from a position closer to the display to position farther from the display, the hand remaining in the engagement volume 910. The use of 3-D imaging is described in greater detail hereinbelow. Similarly, the user can cause the selected window to be grabbed and moved by bringing fingers together with their thumb, and subsequently moving the hand. The selected window will move in relation to the user hand movement until the hand is opened up to release the selected window. Additional control over the selected window can be defined in response to particular hand movements or hand gestures. In accordance with another aspect of the present invention, the selected window will move in response to the user pointing their hand, thumb, or finger in a particular direction. For example, if the user points their index finger to right, the window will move to the right within the display. Similarly, if the user points to the left, up, or down the selected window will move to the left, up or down within the display, respectively. Additional window controls can be achieved through the use of similar hand gestures or motions.

In accordance with another aspect of the invention, the system is configurable such that an individual user selects the particular hand gestures that they wish to associate with particular commands. The system provides default settings that map a given set of gestures to a given set of commands. This mapping, however, is configurable such that the specific command executed in response to each particular hand gesture is definable by each user. For example, one user may wish to point directly at the screen with their

index finger to grab the selected window for movement while another user may wish to bring their fingers together with their thumb to grab the selected window. Similarly, one user may wish to point a group of fingers up or down in order to move a selected window up or down, while another user may wish to open the palm of their hand toward the cameras and then move their opened hand up or down to move a selected window up or down. All given gestures and commands are configurable by the individual users to best suit that particular user's individual personal preferences.

Similarly, in accordance with another aspect of the present invention, the system may include a "Record and Define Gesture" mode. In the "Record and Define Gesture" mode, the system records hand gestures performed by the user. The recorded gestures are then stored in the system memory to be recognized during normal operation. The given hand gestures are then associated with a particular command to be performed by the system in response to that particular hand gesture. With such capability, a user may further tailor the system to their personal preference or, similarly, may tailor system operation to respond to specific commands most appropriate for particular applications.

In a similar fashion, the user can choose the particular words, from a given set, they wish to use for a particular command. For example, one user may choose to say "Release" to stop moving a window while another may wish to say, "Quit". This capability allows different users, which may prefer to use different words for a given command, the ability to tailor the system in a way most efficient for their personal use.

The present invention can be utilized in an expansive list of applications. The following discussion is exemplary of only a few applications with which the present invention may be utilized. One such application is associated with user control of a presentation, or similar type of briefing application, wherein the user makes a presentation on a projection type screen to a group of listeners.

Referring now to FIG. 9B, there is illustrated a multiple monitor implementation. Here, the system includes three monitors (or displays) through which the user 900 exercises control of GUI features; a first display 912, a second display 914, and a third display 916. The cameras (906 and 908) are similarly situated as in FIG. 9A, to define the engagement volume 910. By utilizing the "flick" or "bump" motion(s) as performed by a hand 918 of the user 900, the user 900 can move a window 920 from the first display

912 to the second display 914, and further from the second display 914 to the third display 916. The flick motion of the user hand 918 can effect movement of the window 920 from the first display 912 to the third display 916 in a single window movement, or in multiple steps through the displays (914 and 916) using corresponding multiple hand motions. Of course, control by the user 900 occurs only when the user hand 918 breaks the engagement plane 912, and is determined to be a control object (*i.e.*, an object meeting parameters sufficient to effect control of the computer).

As mentioned hereinabove, the user 900 is located generally in front of the displays (912, 914, and 916), which is also within the imaging capabilities of the pair of cameras (906 and 908). The microphone 904 is suitably located to receive user voice signals. The cameras (906 and 908) are mounted substantially parallel to each other and on a horizontal plane above the displays (912, 914, and 916), and separated by a distance that provides optimum detection and tracking for the given cameras and the engagement volume 910.

In operation, the user 900 moves the hand 918 through the engagement plane 912 and into the engagement volume 910 established for the system. The system, which had detected and tracked the hand 918 before it entered the volume 912, begins providing feedback to the user 900 by displaying the hand shaped icon 922 on one of the displays (912, 914, and 916). The microphone 904 provides additional means for the user 900 to invoke one or more verbal commands in order to act upon the selected window 920 under the corresponding icon 922. The window 920 directly underneath the hand shaped icon is the selected window. When the user hand 918 enters the volume 910, it is recognized as a control object. The corresponding icon 922 is presented by the system on the computer display 912. By dwelling a predetermined amount of time, the associated window is assigned for control. The user 900 causes the selected window to move up or down within the display by invoking the 'Move' command as explained above and then moving the hand up or down, or to move across one or more of the monitors (914 and 916) by invoking the 'Flick' command and then using the flick hand motion. Of course, if the second display 914 was the initial point of control, the user 900 can cause the window 920 to be moved left to the first display 912, or right to the third display 916.

Movement of the window is terminated (or "released") when the user hand dwells for a time longer than a predetermined dwell time, or out of the engagement volume 910.

Alternatively, the user may speak the word "Release" and the system will stop moving the selected window in response to the user's hand motion. Release may also be accomplished by dwelling a bit while in Move, and/or Scroll modes. The user may also act upon a selected window with other actions. By speaking the words, "Close", "Minimize", or "Maximize" the selected window is respectively closed, minimized or maximized. By speaking the word "Raise", the selected window is brought to the foreground, and by speaking "Send to Back", the selected window is sent behind (to the background) all other open windows. By speaking "Scroll", the user initiates a scrolling mode on the selected window. The user may control the rate of the scroll by the position of the hand. The hand shaped icon tracks the user's hand position, and the rate of the scrolling of the selected window is proportional to the distance between the current hand icon position and the position of the hand icon at the time the scrolling is initiated. Scrolling can be terminated by the user speaking "Release" or by the user moving their hand behind the engagement plane and out of the engagement volume. These are just a few examples of the voice recognition perceptual computer control capabilities of the disclosed architecture.

Referring now to FIG. 10, there is illustrated a briefing room environment where voice and/or gestures are utilized to control a screen projector via a computer system configured in accordance with an aspect of the present invention. The briefing room 1000 comprises a large briefing table 1002 surrounded on three sides by numerous chairs 1004, a computer 1006, a video projector 1008, and a projector screen 1010. Utilization of the present invention adds additional elements comprising the disclosed perceptual software 1012, two video cameras (1014 and 1016) and a microphone 1018. In this application, a user 1020 is positioned between the projector screen 1010 and briefing table 1002 at which the audience is seated. A top face 1022 of an engagement volume 1024 is defined by rectangular area 1026. Similarly, a front surface indicated at 1028 represents an engagement plane.

As the user gives the presentation, the user controls the content displayed on the projection screen 1010 and advancement of the slides (or presentation images) by moving

their hand(s) through the engagement plane 1028 into the engagement volume 1024, and/or speaking commands recognizable by the system. Once inside the engagement volume 1024, a simple gesture is made to advance to the next slide, back-up to a previous slide, initiate an embedded video, or to effect one of a number of many other presentation capabilities.

A similar capability can be implemented for a home media center wherein the user can change selected video sources, change channels, control volume, advance chapter and other similar functions by moving their hand across an engagement plane into an engagement volume and subsequently performing the appropriate hand gesture. Additional applications include perceptual interfaces for TabletPCs, Media center PCs, kiosks, hand held computers, home appliances, video games, and wall sized displays, along with many others.

It is appreciated that in more robust implementations, instead of the engagement volume being fixed at a position associated with the location of the cameras that requires the presenter to operate according to the location of the engagement volume, the system can be configured such that the engagement volume travels with the user (in a "roaming" mode) as the user moves about the room. Thus, the cameras would be mounted on a platform that rotates such that the rotation maintains the cameras substantially equidistant from the presenter. The presenter may carry a sensor (*e.g.*, an RFID tag) that allows the system to sense or track the general location of the presenter. The system would then affect rotation of the camera mount to "point" the cameras at the presenter. In response thereto, the engagement volume may be extended to the presenter allowing control of the computer system as the presenter moves about. The process of "extending" the engagement volume can include increasing the depth of the volume such that the engagement plane surface moves to the presenter, or by maintaining the volume dimensions, but moving the fixed volume to the presenter. This would require on-the-fly focal adjustment of the cameras to track quick movements in the depth of objects in the volume, but also the movement of the presenter.

Another method of triggering system attention in this mode would be to execute a predefined gesture that is not likely to be made unintentionally, *e.g.*, raising a hand.

It is also appreciated that the system is configurable for individual preferences such that the engagement volume of a first user may be different from the volume of a second user. For example, in accordance with a user login, or other unique user information, the user preferences may be retrieved and implemented automatically by the system. This can include automatically elevating the mounted cameras for a taller person by using a telescoping camera stand so that the cameras are at the appropriate height of the particular user, whether sitting or standing. This also includes, but is not limited to, setting the system for "roaming" mode.

Referring now to FIG. 11, there is illustrated a block diagram of a computer operable to execute the present invention. In order to provide additional context for various aspects of the present invention, FIG. 11 and the following discussion are intended to provide a brief, general description of a suitable computing environment 1100 in which the various aspects of the present invention may be implemented. While the invention has been described above in the general context of computer-executable instructions that may run on one or more computers, those skilled in the art will recognize that the invention also may be implemented in combination with other program modules and/or as a combination of hardware and software. Generally, program modules include routines, programs, components, data structures, etc., that perform particular tasks or implement particular abstract data types. Moreover, those skilled in the art will appreciate that the inventive methods may be practiced with other computer system configurations, including single-processor or multiprocessor computer systems, minicomputers, mainframe computers, as well as personal computers, hand-held computing devices, microprocessor-based or programmable consumer electronics, and the like, each of which may be operatively coupled to one or more associated devices. The illustrated aspects of the invention may also be practiced in distributed computing environments where certain tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, program modules may be located in both local and remote memory storage devices.

A computer typically includes a variety of computer readable media. Computer readable media can be any available media that can be accessed by the computer and includes both volatile and nonvolatile media, removable and non-removable media. By

way of example, and not limitation, computer readable media can comprise computer storage media and communication media. Computer storage media includes both volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information such as computer readable instructions, data structures, program modules or other data. Computer storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital video disk (DVD) or other optical disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to store the desired information and which can be accessed by the computer. Communication media typically embodies computer readable instructions, data structures, program modules or other data in a modulated data signal such as a carrier wave or other transport mechanism and includes any information delivery media. The term "modulated data signal" means a signal that has one or more of its characteristics set or changed in such a manner as to encode information in the signal. By way of example, and not limitation, communication media includes wired media such as a wired network or direct-wired connection, and wireless media such as acoustic, RF, infrared and other wireless media. Combinations of the any of the above should also be included within the scope of computer readable media.

With reference again to FIG. 11, the exemplary environment 1100 for implementing various aspects of the invention includes a computer 1102, the computer 1102 including a processing unit 1104, a system memory 1106, and a system bus 1108. The system bus 1108 couples system components including, but not limited to the system memory 1106 to the processing unit 1104. The processing unit 1104 may be any of various commercially available processors. Dual microprocessors and other multi-processor architectures also can be employed as the processing unit 1104.

The system bus 1108 can be any of several types of bus structure including a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of commercially available bus architectures. The system memory 1106 includes read only memory (ROM) 1110 and random access memory (RAM) 1112. A basic input/output system (BIOS), containing the basic routines that help to transfer

information between elements within the computer 1102, such as during start-up, is stored in the ROM 1110.

The computer 1102 further includes a hard disk drive 1114, a magnetic disk drive 1116, (*e.g.*, to read from or write to a removable disk 1118) and an optical disk drive 1120, (*e.g.*, reading a CD-ROM disk 1122 or to read from or write to other optical media). The hard disk drive 1114, magnetic disk drive 1116 and optical disk drive 1120 can be connected to the system bus 1108 by a hard disk drive interface 1124, a magnetic disk drive interface 1126 and an optical drive interface 1128, respectively. The drives and their associated computer-readable media provide nonvolatile storage of data, data structures, computer-executable instructions, and so forth. For the computer 1102, the drives and media accommodate the storage of broadcast programming in a suitable digital format. Although the description of computer-readable media above refers to a hard disk, a removable magnetic disk and a CD, it should be appreciated by those skilled in the art that other types of media which are readable by a computer, such as zip drives, magnetic cassettes, flash memory cards, digital video disks, cartridges, and the like, may also be used in the exemplary operating environment, and further that any such media may contain computer-executable instructions for performing the methods of the present invention.

A number of program modules can be stored in the drives and RAM 1112, including an operating system 1130, one or more application programs 1132, other program modules 1134 and program data 1136. It is appreciated that the present invention can be implemented with various commercially available operating systems or combinations of operating systems.

A user can enter commands and information into the computer 1102 through a keyboard 1138 and a pointing device, such as a mouse 1140. Other input devices (not shown) may include one or more video cameras, one or microphones, an IR remote control, a joystick, a game pad, a satellite dish, a scanner, or the like. These and other input devices are often connected to the processing unit 1104 through a serial port interface 1142 that is coupled to the system bus 1108, but may be connected by other interfaces, such as a parallel port, a game port, an IEEE 1394 serial port, a universal serial bus ("USB"), an IR interface, etc. A monitor 1144 or other type of display device is also

connected to the system bus 1108 *via* an interface, such as a video adapter 1146. In addition to the monitor 1144, a computer typically includes other peripheral output devices (not shown), such as speakers, printers etc.

The computer 1102 may operate in a networked environment using logical connections to one or more remote computers, such as a remote computer(s) 1148. The remote computer(s) 1148 may be a workstation, a server computer, a router, a personal computer, portable computer, microprocessor-based entertainment appliance, a peer device or other common network node, and typically includes many or all of the elements described relative to the computer 1102, although, for purposes of brevity, only a memory storage device 1150 is illustrated. The logical connections depicted include a LAN 1152 and a WAN 1154. Such networking environments are commonplace in offices, enterprise-wide computer networks, intranets and the Internet.

When used in a LAN networking environment, the computer 1102 is connected to the local network 1152 through a network interface or adapter 1156. When used in a WAN networking environment, the computer 1102 typically includes a modem 1158, or is connected to a communications server on the LAN, or has other means for establishing communications over the WAN 1154, such as the Internet. The modem 1158, which may be internal or external, is connected to the system bus 1108 *via* the serial port interface 1142. In a networked environment, program modules depicted relative to the computer 1102, or portions thereof, may be stored in the remote memory storage device 1150. It will be appreciated that the network connections shown are exemplary and other means of establishing a communications link between the computers may be used.

Further, a camera 1160 (such as a digital/electronic still or video camera, or film/photographic scanner) capable of capturing a sequence of images 1162 can also be included as an input device to the computer 1102. While just one camera 1160 is depicted, multiple cameras 1160 could be included as input devices to the computer 1102. The images 1162 from the one or more cameras 1160 are input into the computer 1102 *via* an appropriate camera interface 1164. This interface 1164 is connected to the system bus 1108, thereby allowing the images 1162 to be routed to and stored in the RAM 1112, or one of the other data storage devices associated with the computer 1102. However, it is noted that image data can be input into the computer 1102 from any of the

aforementioned computer-readable media as well, without requiring the use of the camera 1160.

Referring now to FIG. 12, there is illustrated a network implementation 1200 of the present invention. The implementation 1200 includes a first perceptual system 1202 and a second perceptual system 1204, both operational according to the disclosed invention. The first system 1202 includes cameras 1206 (also denoted C1 and C2) mounted on a rotational and telescoping camera mount 1208. A first user 1210 located generally in front of the first system 1202 effects control of a GUI Content A of the first system 1202 in accordance with the novel aspects of the present invention by introducing hand gestures into an engagement volume 1211 and/or voice signals *via* a microphone. The first user 1210 may roam about in front of the cameras 1206 in accordance with the "roaming" operational mode described previously, or may be seated in front of the cameras 1206. The second system 1204 includes cameras 1212 (also denoted C3 and C4) mounted on a rotational and telescoping camera mount 1214. A second user 1216 located generally in front of the second system 1204 effects control of a GUI Content B of the second system 1204 in accordance with the novel aspects of the present invention by introducing hand gestures into an engagement volume 1217 and/or voice signals using a microphone. The second user 1216 may roam about in front of the cameras 1212 in accordance with the "roaming" operational mode described previously, or may be seated in front of the cameras 1212.

The first and second systems (1202 and 1204) may be networked in a conventional wired or wireless network 1207 peer configuration (or bus configuration by using a hub 1215). This particular system 1200 is employed to present both Content A and Content B *via* a single large monitor or display 1218. Thus the monitor 1218 can be driven by either of the systems (1202 and 1204), as can be provided by conventional dual-output video graphics cards, or the separate video information may be transmitted to a third monitor control system 1220 to present the content. Such an implementation finds application where a side-by-side comparison of product features is being presented, or other similar applications where two or more users may desire to interact. Thus, Content A and Content B can be presented on a split screen layout of the monitor 1218. Either or

both users (1210 and 1216) can also provide keyboard and/or mouse input to facilitate control according to the present invention.

3-D IMAGING IMPLEMENTATIONS

Referring now to FIG. 13, there is illustrated a medical operating room system 1300 that uses the engagement volume in accordance with the present invention. An operating room 1302 includes an operating table 1304 on which a patient 1305 is placed. A doctor (or medical person) 1306 is positioned to one side of the table 1304 in order to effectively operate on the patient 1305. However, it is to be appreciated that the medical person 1306 may be required to move around the table 1304 and operate from various positions and angles.

The operating room system 1300 also includes an operation computer system 1308 used by the medical person 1306 to facilitate the operation. In this particular embodiment, the operation computer system 1308 comprises three computer systems: a first computer system 1310, a second computer system 1312, and a third computer system 1314. The first system 1310 includes a first monitor (or display) 1316, the second system 1312 includes a second display 1318, and the third system 1314 includes a third display 1320. Medical information related to the patient 1305 can be displayed on the any one or more of the monitors (1316, 1318 and 1320) before, during, and/or after the operation. Note that the computer and displays can be oriented or positioned in any manner suitable for easy use and viewing by operating room personnel.

The operation computing system 1308 also includes at least a pair of cameras 1322 suitably designed for capturing images of at least the hands, arms, head, and general upper torso appendage positions, to the level of hand and finger positions of the medical person 1306. The cameras 1322 can be connected to a single computer system for the input of image data, and thereafter, the image data distributed among the computing systems (1310, 1312, and 1314) for processing. The three computer systems (1310, 1312, and 1314) are networked on a wired network 1324, which network 1324 can connect to a larger hospital or facility-wide network, for example. Note that it is not required to have three computer systems. Alternatively, in such environments where the network 1324 can present a bottleneck to such data transfers, a gigabit or faster network

can be employed internally and locally for high-speed communication of the image data between the computer systems (1310, 1312, and 1314) or to a fourth computer system (not shown) on the local high-speed network that can more efficiently and quickly process and present the image data to any one or more of the displays (1316, 1318, and 1320). The disclosed invention is not restricted to more computers or fewer computers. This is to indicate that the system can employ a plurality of computers for presenting the same information from multiple perspectives (as could be beneficial in an operating room environment), and different information from each system, for example.

In one implementation, the operation computing system 1308 develops an engagement volume 1326 above the operating table 1304, which volume envelops part or all the operation area of patient 1305. Thus, the table 1304, patient 1305, and volume 1326 are all at a height suitable for operation such that the hands of the medical person 1306 can engage the volume 1326 at an appropriate height to be detected and tracked by the computing system 1308. Hand gestures of the medical person 1306 are then imaged, tracked, and processed, as described hereinabove, and more specifically, with respect to FIG. 9, to facilitate controlling the presentation of information on one or more of the displays (1316, 1318, and 1320) *via* associated computing systems, as this can also entail audio I/O. In support of voice commands, the medical person 1306 can be outfitted with a wireless portable microphone system 1328 that includes a power supply, microphone, and transmitter for communicating wirelessly with a computer wireless transceiver system 1330 of the operation computer system 1308. Thus, voice commands alone or in combination with hand gestures can be used to facilitate the operation.

Referring now to FIG. 14, there is illustrated a medical operating room environment 1400 in which a computer control system 1404 with a wireless control device 1404 is employed. The system 1404 also includes the use of the wireless control device 1404 for control thereof. Here, the engagement volume of FIG. 13 is no longer required or used only marginally. Continuing with the operating room implementation, the medical person 1306 uses the wireless remote control user interface (UI) device 1404 (hereinafter referred to as a "wand") to facilitate control of the operation computer system 1308. The wand 1404 can be positioned on a headpiece 1406 worn by the medical person 1306 to provide the free use of hands during the procedure. The wand

1404 is oriented in parallel with the line of sight (or also called, "gaze") of the person 1306 such that when the person's line of sight is to the system 1308, this is detected as an interaction to be processes by the system 1308. All the person 1306 needs to do is perform head movements to facilitate control of the operation computing system 1308. The wand 1404 includes one or more sensors the outputs of which are transmitted to the transceiver system 1330 and forwarded to the operation computing system 1308 for processing. The wand 1404 and associated computing system and imaging capabilities are described in the following pending U.S. Patent Applications: Serial No. 10/160,692, entitled "A SYSTEM AND PROCESS FOR SELECTING OBJECTS IN A UBIQUITOUS COMPUTING ENVIRONMENT," filed May 31, 2002, and Serial No. 10/160,659, entitled "A SYSTEM AND PROCESS FOR CONTROLLING ELECTRONIC COMPONENTS IN A UBIQUITOUS COMPUTING ENVIRONMENT USING MULTIMODAL INTEGRATION," filed May 31, 2002, both of which are hereby incorporated by reference.

In general, the system 1402 includes the aforementioned wand 1404 in the form of the wireless radio frequency (RF) pointer, which includes an RF transceiver and various orientation sensors. The outputs of the sensors are periodically packaged as orientation signals and transmitted using the RF transceiver to the computer transceiver 1330, which also has a RF transceiver to receive the orientation messages transmitted by the wand 1404. The orientation signals of the wand 1404 are forwarded to the computer system 1308. The computer system 1308 is employed to compute the orientation and location of the wand 1404 using the orientation signals, as are images of the wand 1404 captured by the cameras 1322. The orientation and location of the wand 1404 is in turn used to determine if the wand 1404 is being pointed at an object in the operating room environment 1400 that is controllable by the computer system 1308 *via* the network 1324, such as one of the displays (1316, 1318, or 1320). If so, the object is selected.

The wand 1404 specifically includes a case having a shape with a defined pointing end, a microcontroller, the aforementioned RF transceiver and orientation sensors which are connected to the microcontroller, and a power supply (*e.g.*, batteries) for powering these electronic components. The orientation sensors of the wand 1404 include at least, an accelerometer, which provides separate x-axis and y-axis orientation

signals, and a magnetometer, which provides separate tri-axial measurements (x-axis, y-axis, and z-axis) orientation signals. These electronics are housed in a case that resembles a handheld wand. However, the packaging can be of any form factor such that the functionality of the wand 1404 can be used for the particular purpose.

As indicated previously, the orientation signals generated by the wand 1404 include the outputs of the sensors. To this end, the wand microcontroller periodically reads and stores the outputs of the orientation sensors. Whenever a request for an orientation signal is received (or it is time to generate such a signal if the pointer is programmed to do so without a request), the microcontroller includes the last-read outputs from the accelerometer and magnetometer in the orientation signal.

The wand 1404 also includes other electronic components such as a user activated switch or button, and a series of light emitting diodes (LEDs). The user-activated switch, which is also connected to the microcontroller, is employed for the purpose of instructing the computer to implement a particular function, such as will be described later. To this end, the state of the switch in regard to whether it is activated or deactivated at the time an orientation message is packaged is included in that message for transmission to the computer. The series of LEDs includes a pair of differently colored, visible spectrum LEDs, which are connected to the microcontroller, and which are visible from the outside of the pointer's case when lit. These LEDs are used to provide status or feedback information to the user, and are controlled via instructions transmitted to the pointer by the computer.

However, as will be described in greater detail hereinbelow, since the wand 1404 includes at least one motion sensor, the user activated switch can be implemented in an alternative manner using hands-free control thereof *via* head movements, for example, or a combination of voice activation, and/or head movement, just to name a few.

The foregoing system 1402 is utilized to select an object by having the user simply point to the object or feature with the wand 1404. This entails the computer system 1308 first receiving the orientation signals transmitted by the wand 1404. For each message received, the computer 1308 derives the orientation of the wand 1404 in relation to a predefined coordinate system of the environment in which the wand 1404 is operating using the orientation sensor readings contained in the message. In addition, the

video output from the video cameras 1322 is used to ascertain the location of the wand 1404 at a time substantially contemporaneous with the generation of the orientation signals and in terms of the predefined coordinate system. Once the orientation and location of the wand 1404 are computed, they are used to determine whether the wand 1404 is being pointed at an object in the environment that is controllable by the computer system 1308. If so, then that object is selected for future control actions.

The computer system 1308 derives the orientation of the wand 1404 from the orientation sensor readings contained in the orientation signals, as follows. First, the accelerometer and magnetometer output values contained in the orientation signals are normalized. Angles defining the pitch of the wand 1404 about the x-axis and the roll of the device about the y-axis are computed from the normalized outputs of the accelerometer. The normalized magnetometer output values are then refined using these pitch and roll angles. Next, previously established correction factors for each axis of the magnetometer, which relate the magnetometer outputs to the predefined coordinate system of the environment, are applied to the associated refined and normalized outputs of the magnetometer. The yaw angle of the wand 1404 about the z-axis is computed using the refined magnetometer output values. The computed pitch, roll and yaw angles are then tentatively designated as defining the orientation of the wand 1404 at the time the orientation signals are generated.

It is next determined whether the wand 1404 is in a right-side up or up-side down position at the time the orientation signals were generated. If the wand 1404 was in the right-side up position, the previously computed pitch, roll and yaw angles are designated as the defining the finalized orientation of the wand 1404. However, if it is determined that the wand 1404 was in the up-side down position at the time the orientation message was generated, the tentatively designated roll angle is corrected accordingly, and then the pitch, yaw and modified roll angle are designated as defining the finalized orientation of the wand 1404.

In the foregoing description, it is assumed that the accelerometer and magnetometer of the wand 1404 are oriented such that their respective first axis corresponds to the x-axis which is directed laterally to a pointing axis of the wand 1404, and their respective second axis corresponds to the y-axis, which is directed along the

pointing axis of the wand 1404, and the third axis of the magnetometer corresponds to the z-axis, which is directed vertically upward when the wand 1404 is positioned right-side up with the x and y axes lying in a horizontal plane.

The computer system 1308 derives the location of the wand 1404 from the video output of the video cameras 1322, as follows. In the wand 1404, there is an infrared (IR) LED connected to a microcontroller that is able to emit IR light outside the wand 1404 case when lit. The microcontroller causes the IR LEDs to flash. In addition, the aforementioned pair of digital video cameras 1322 each have an IR pass filter that results in the video image frames capturing only IR light emitted or reflected in the environment toward the cameras 1322, including the flashing from the wand 1404 IR LED which appears as a bright spot in the video image frames. The microcontroller causes the IR LED to flash at a prescribed rate that is approximately one-half the frame rate of the video cameras 1322. This results in only one of each pair of image frames produced by a camera having the IR LED flashes depicted in it. This allows each pair of frames produced by a camera to be subtracted to produce a difference image, which depicts for the most part only the IR emissions and reflections directed toward the camera which appear in one or the other of the pair of frames but not both (such as the flash from the IR LED of the pointing device). In this way, the background IR in the environment is attenuated and the IR flash becomes the predominant feature in the difference image.

The image coordinates of the pixel in the difference image that exhibits the highest intensity is then identified using a standard peak detection procedure. A conventional stereo image technique is employed to compute the 3-D coordinates of the flash for each set of approximately contemporaneous pairs of image frames generated by the pair of cameras 1322 using the image coordinates of the flash from the associated difference images and predetermined intrinsic and extrinsic camera parameters. These coordinates represent the location of the wand 1404 (as represented by the location of the IR LED) at the time the video image frames used to compute the coordinates were generated by the cameras 1322.

The orientation and location of the wand 1404 at any given time is used to determine whether the wand 1404 is being pointed at an object in the environment that is controllable by the computer system 1308. In order to do so, the computer system 1308

must know what objects are controllable and where they exist in the environment. This requires a model of the environment. In the present system and process, the location and extent of objects within the environment that are controllable by the computer system 1308 are modeled using 3-D Gaussian blobs defined by a location of the mean of the blob in terms of its environmental coordinates and a covariance.

At least two different methods have been developed to model objects in the environment. The first method involves the user inputting information identifying the object that is to be modeled. The user then activates the switch on the pointing device and traces the outline of the object. Meanwhile, the computer system 1308 is running a target training procedure that causes requests for orientation signals to be sent to the wand 1404 at a prescribed request rate. The orientation signals are input when received, and for each orientation signal, it is determined whether the switch state indicator included in the orientation signal indicates that the switch is activated. Whenever it is initially determined that the switch is not activated, the switch state determination action is repeated for each subsequent orientation signal received until an orientation signal is received that indicates the switch is activated. At that point, each time it is determined that the switch is activated, the location of the wand 1404 is ascertained, as described previously, using the digital video input from the pair of video cameras 1322. When the user is done tracing the outline of the object being modeled, he or she deactivates the switch. The target training (or calibration) process detects this as the switch having been deactivated after first having been activated in the immediately preceding orientation signal. Whenever such a condition occurs, the tracing procedure is deemed to be complete and a 3-D Gaussian blob representing the object is established using the previously ascertained wand locations stored during the tracing procedure.

The second method of modeling objects during a calibration process once again begins by the user inputting information identifying the object that is to be modeled. However, in this case the user repeatedly points the wand 1404 at the object and momentarily activates the switch on the wand 1404, each time pointing the wand 1404 from a different location within the environment. Meanwhile, the computer system 1308 is running a target training algorithm that causes requests for orientation signals to be sent

to the wand 1404 at a prescribed request rate. Each orientation message received from the wand 1404 is input until the user indicates the target training inputs are complete.

For each orientation signal input, it is determined whether the switch state indicator contained therein indicates that the switch is activated. Whenever it is determined that the switch is activated, the orientation of the wand 1404 is computed, as described previously, using orientation sensor readings also included in the orientation message. In addition, the location of the wand 1404 is ascertained using the inputted digital video from the pair of video cameras 1322. The computed orientation and location values are stored.

Once the user indicates the target training inputs are complete, the location of the mean of a 3-D Gaussian blob that will be used to represent the object being modeled is computed from the stored orientation and location values of the wand 1404. The covariance of the Gaussian blob is then obtained in one of various ways. For example, it can be a prescribed covariance, a user input covariance, or the covariance can be computed by adding a minimum covariance to the spread of the intersection points of rays defined by the stored orientation and location values of the wand 1404.

With a Gaussian blob model of the environment in place, the orientation and location of the wand 1404 is used to determine whether the wand 1404 is being pointed at an object in the environment that is controllable by the computer system 1308. In one version of this procedure, for each Gaussian blob in the model, the blob is projected onto a plane that is normal to either a line extending from the location of the wand 1404 to the mean of the blob, or a ray originating at the location of the wand 1404 and extending in a direction defined by the orientation of the wand 1404. The value of the resulting projected Gaussian blob at a point where the ray intersects the plane, is computed. This value represents the probability that the wand 1404 is pointing at the object associated with the blob under consideration.

Next, the probability is computed that represents the largest value computed for the Gaussian blobs, if any. At this point, the object associated with the Gaussian blob from which the largest probability value was derived could be designated as being the object at which the wand 1404 is pointing. However, an alternative thresholding procedure could be employed instead. In this alternate version, it is first determined

whether the largest probability value exceeds a prescribed minimum probability threshold. Only if the threshold is exceeded is the object associated with the projected Gaussian blob from which the largest probability value was derived designated as being the object at which the wand 1404 is pointing. The minimum probability threshold is chosen to ensure the user is actually pointing at the object and not just near the object without an intent to select it.

In an alternate procedure for determining whether the wand 1404 is being pointed at an object in the environment 1400 that is controllable by the computer system 1308, for each Gaussian blob, it is determined whether a ray originating at the location of the wand 1404 and extending in a direction defined by the orientation of the wand 1404 intersects the blob. Next, for each Gaussian blob intersected by the ray, it is determined what the value of the Gaussian blob is at a point along the ray nearest the location of the mean of the blob. This value represents the probability that the wand 1404 is pointing at the object associated with the Gaussian blob. The rest of the procedure is similar to the first method, in that, the object associated with the Gaussian blob from which the largest probability value was derived could be designated as being the object at which the wand 1404 is pointing. Alternatively, it is first determined whether the probability value identified as the largest exceeds the prescribed minimum probability threshold. If the threshold is exceeded, only then is the object associated with the projected Gaussian blob from which the largest probability value was derived designated as being the object at which the wand 1404 is pointing.

Hands-free control of the operation computing system 1308 using the head mounted wand 1404 involves generating at least a series of calibrated head movements. Moreover, since the person 1306 also uses the wireless microphone system 1328, voice commands can be employed alone or in combination with the head movements to enhance control of the computer system 1308. With the implementation of one or more motion sensors therein, *e.g.*, accelerometers, velocity and/or acceleration data can be measured and resolved as the "switch" signal of the wand 1404 to initiate or terminate an action without physically having to move a switch with a finger, which would be extremely cumbersome and risky (insofar at least as sterilization and the transmission of germs is involved) in an operating room environment. For example, when the system

1308 determines that the gaze of the medical person 1306 is at the second display 1318, a simple left-right head movement can be interpreted to initiate a paging action such that displayed images are changed similar to a person thumbing through pages of a book. Thereafter, an up-down head nod could be used to stop the paging process. Alternatively, the paging process could be initiated by voice command after the system 1308 ascertains that the gaze is directed at the second display 1318.

If more than one wand 1404 was employed by operating room personnel, the wands can be uniquely identified by an RF tagging system, such that signals transmitted to the computer system 1308 are interpreted in association with different personnel. For example, the doctor in charge of the operation and his or her assisting nurse could each have a head mounted wand. The system 1308 can be suitably designed to discriminate the wand signals according to a unique tag ID that accompanies each signal transmitted to the computer system 1308. Such tagging system can also be used as a method of prioritizing signals for controlling the computer. For example, the system can be configured to prioritize signals received from the doctor over those signals received from the assisting nurse.

In a more sophisticated implementation, the computer system 1308 employs the classifier system described hereinabove to learn the movements of personnel over time. For example, body movements of one person typically differ from the way a body movement of another may be used to control the system 1308. Thus, instead of the user of the wand 1404 conforming to rigid criteria of signaling required by the computer system algorithm, the system 1308 can employ the classifier to learn the particular movements of a given user. Once the user "logs in" to the system 1308, these customized movement signals (and voice signals, for example) can then be activated for use by the system 1308 for that user.

It is to be appreciated that once the use of a remote wireless system 1404 is employed, other internal and external signals can be input thereto for transmission to and control of the system 1308. For example, the heart rate of the person 1306 can be monitored and input to the wand system 1404 or wireless voice system 1328 for wireless input to the system 1308 to monitor the state of the person 1306. If, during a lengthy operation, the system 1308 detects that the physical condition of the person 1306 is

deteriorating, the classifier can be used to modify how the movement and voice signals are processed for controlling the system 1308. A faster heart rate can indicate faster speech and/or head movements that would then be compensated for in the system 1308 using the classifier. Of course, these parameters would be determined on a user-by-user basis.

In accordance with the orientation signals received from the wand 1404, the system 1308 can determine a number of factors about the person 1306. The system 1308 can determine when the person 1306 (or what person(s)) is looking at the system 1308. For example, if the orientation of the wand 1404 indicates that the head position (or gaze) of the person 1306 matches orientation associated with looking at any of the three monitors (1316, 1318, or 1320), here, the second monitor 1318, the system 1308 then responds according to signals received thereafter, until the viewing session associated with the second monitor 1318 is terminated by voice and/or head movements.

Where only one wand 1404 is provided, the system 1308 can re-associate the wand 1404 with a user profile of another person 1306 who will use the wand 1404. There exists a database of user profiles and tag associations such that invocation of the wand tag (or ID) with the user log-in name automatically executes the user profile for use with the wand 1404. This way, individualized user commands in the form of head movements, voice commands, etc., are automatically invoked at the user log-in process.

The system 1308 can also employ a bi-directional activation scheme wherein the user initiates a user command for starting a session, and the system 1308 responds with a signal that further requires a user response to confirm that a session is to begin. For example, the person 1306 can initiate a session by motioning an up-down head nod repeatedly for three cycles. The system 1308 receives the corresponding three cycles of up-down nod signals that are interpreted to start a session of that person 1306. In order to ensure that the head nod was not made inadvertently, the system 1308 responds by presenting an image on the first display 1316, and at which the person 1306 must point the wand 1404 to confirm the start of the session. Of course, other signals can be used to confirm session start. For example, the user can look to the ceiling, which orientation of the wand 1404 in a substantially vertical direction is interpreted to confirm the start of a session. Obviously, the number and combination of head movements and/or voice

commands that can be employed in the present system are numerous, and can be used in accordance with user preferences.

In the system 1402, the transceiver system 1330 can be used for wireless communication for both the wand system 1404 and voice communications system 1328. Thus, the wand link can be of one frequency, and the voice communication link another frequency. The computer system 1308 is configured to accommodate both by proving frequency discrimination and processing so that signal streams can be filtered and processed to extract the corresponding wand and voice signals.

Referring now to FIG. 15, there is illustrated a flowchart of a process from the perspective of the person for using the system of FIG. 14. At 1500, the user performs a calibration process that comprises associating a number of head movements and/or voice commands with user commands. This also includes using voice commands singly to control the operation computer system or in combination with the head movements to do so. The calibration process can be performed well in advance of use in the operating room, and updated as the user chooses to change movements and/of voice signals with user commands. At 1502, the person initiates a session with the computer system using one or more user commands. At 1504, the person then inputs one or more of the user commands to control the computing system. At 1506, the person terminates the session using one or more of the user commands. The process then reaches a Stop block.

Referring now to FIG. 16, there is illustrated a flowchart of a process from the perspective of the system of FIG. 14. At 1600, the calibration process occurs where the system associates wand device signals and/or voice signals with user commands specified by the person. The calibration process ends. At 1602, the user wand signals are received and processed by the system. At 1604, the system determines if the processed user command(s) indicate that a session is to be started with that user. If NO, flow is back to the input of 1602 to continue to processing received device and voice signals. If YES, flow is to 1606 to identify the user. This can occur by the system processing the received signal and extracting the device tag ID. Prior to use, the tag ID of the wand is programmed for association with a given user. At 1608, the user profile of calibration data is activated for use. Of course, on any given operation, the operating staff are

assigned such that the log-in names of the doctors and assistants can be entered prior to beginning the operation. Thus, the user profiles are already activated for processing.

At 1610, the device signals are received and processed to determine the tag ID of the device and to process the user command(s) against the associated profile information to enable the related command. At 1612, where a classifier is employed, the classifier tracks, processes, compares, and updates the user profile when wand movements associated with the particular user command are changed within certain criteria. At 1614, the computer system determines if the session has completed. If NO, flow is back to the input of 1610 to continue to process user commands. If YES, flow is to 1616 to terminate the user session. The process then reaches a Stop block. Of course, from 1616, flow can be brought back to the input of 1602 to continue to process signals from other devices or to prepare for another session, which could occur many times during the operating room event.

Referring now to FIG. 17, there is illustrated a medical environment 1700 in which a 3-D imaging computer control system 1702 is employed to process hand (or body) gestures in accordance with the present invention. The operation computing system 1308 provides 3-D image recognition and processing capability such that the engagement volume of FIG. 13 and the wand 1404 of FIG. 14 are no longer required. The system 1702 can be augmented with voice commands in a manner similar to that described above; however, this is not needed. Audio-visual co-analysis can be used to improve continuous gesture recognition. Here, the transceiver system 1330 is used only for wireless voice communication, when vocalization is employed. For example, the medical person 1306 can simply use the system 1308 as a dictaphone to record voice signals during the operation.

The foregoing system 1702 is used to select an object under computer control in the environment by the computer system 1308 by having the user simply make one or more hand gestures. Of course, this can be done using both hands, which feature will be described in greater detail hereinbelow. This entails the computer system 1308 capturing imaging information about the hand gesture(s), and for each image or series of images received, the computer system 1308 derives the posture, orientation, and location of the hand, pair of hands, or any combination of one or more hands and any other body part

(*e.g.*, the head) (hereinafter grouped and denoted generally as "gesture characteristics", or where specifically related to a hand, as "hand gesture characteristics", or "hand characteristics") in relation to a predefined coordinate system of the environment in which the gesture is employed. Gesture analysis involves tracking the user's hand(s) in real-time. Hidden Markov Models (HMMs) can be employed for recognition of continuous gesture kinematics. In addition, the video output from the video cameras 1322 is used to ascertain the gesture characteristics at a time substantially contemporaneous with the generation of the gesture and in terms of the predefined coordinate system. Once the gesture characteristics are processed, they are used to determine whether an object in the environment should be controlled by the computer system 1308. If so, then that object is selected for future control actions. Moreover, stochastic tools such as Kalman filtering can be used to predict the position of the hand or hands in subsequent image frames.

Note that an object includes an object or device external to the computer system 1308 and controllable by a wireless and/or wired connection, as well as any internal device or feature that comprises software programs that are used to display images, manipulate data, and move data from one location to another, for example.

The process begins by generating a model of the environment. This process includes, but is not limited to, defining what aspects of the environment will be controlled by the computer system 1308, such as lights, lighting level, room temperature, operating room life support machines and other computer controlled machines in the room, and software controls that will be required or desired of the system 1308 before, during, and/or after the procedure. The software controls comprise the gestures required to initiate image paging, image rotation about a vertex, image rotation about an axis, zooming in and out on an image, providing supplementary data (*e.g.*, video and audio) related to an image being presented or manipulated in a certain way, performing x,y translations of the image, stepped rotation, changing user interface coloring to improve visibility of an image, changing image contrast, changing resolution of an image, playing a series of images quickly or slowly (looping speed), freezing and unfreezing a looping image video (of, for example, echocardiography, transverse CT (Computed Tomography) and cryosection images, CT output, and a fly-through of MRI data), initiating repetitive

image(s) playback (looping), jumping from the first monitor 1316 to another monitor (1318 or 1320), and adjusting audio controls when listening to audio data (*e.g.*, EKG) during the procedure.

The next step is to calibrate the model according to the persons who will be working in the environment and interacting with the system 1308. Unique user profiles can be generated for each person interacting with the system 1308 by employing a tagging system that can discriminate the various users. This can be accomplished in several ways. One method provides a unique RF tag to each user. A triangulation system can be utilized to continually monitor the location of a given user, and associate the location data with the captured image data such that gestures are from that location will be processed against that user profile to properly execute the user command.

Another method is to employ several camera sets, where each set is dedicated to a specific user or location in which the user will work. The user could also be clothed in a colored uniform where the combination of color, gesture, and location uniquely identify the command issued by that user to the system 1308. As mentioned hereinabove, the system 1308 can be programmed to invoke a bi-directional confirmation system such that each user gesture is followed by a confirmation request to ensure that the proper user command is issued. Feedback can be provided by displaying the command in large text or outputting the anticipated command in audio to the user, after which the user responds by voice or with another gesture to accept or reject the command.

The imaging system 1308 detects gesture object (or hand) depth or distance from the system 1308 to facilitate discriminating between a plurality of gesture sources. The gesture sources can include a single hand, two hands, one hand of two people, etc. RF triangulation techniques can be used to accurately determine the gesture source(s). Thus, the gesture source includes an RF tag. If two hands are being used in gesticulation, then each hand can include a unique RF tag. Other depth determination systems can be employed to accurately discriminate the gesture sources, such as infrared.

As with other implementations mention above, the environment needs to be modeled for all objects to be controlled or interacted with, including both hardware and software. The gestures are then defined and associated with the objects. This can further include the use of voice commands, and where the wireless remote device is worn in

alignment with the person's line-of-sight, the additional combination of "gaze" signals, where the gaze signals are defined as those wireless device (or wand) signals generated when the person looks in a direction to effect object interaction.

The system 1308 can also be configured to determine when the operator is generally facing the system 1308. A facial image can be captured and processed with facial features providing generally the data needed to make such an automatic determination. Another method includes wearing a multi-colored uniformed such that one color is associated with the person facing the system 1308, while another imaged color indicates the person is not facing the system 1308. Still another method employs a reflective surface on the front of the person such that the presence of reflective signals indicates the person is facing the system 1308.

The system 1308 is capable of determining when one person programmed to interact therewith has been replaced by another. This causes an automatic change of user profiles to enable the present user's gestures for corresponding user commands and control of the system 1308. Again, this can be facilitated by a color scheme whereby each medical person is uniquely identified in the system 1308 with a unique color. Any sort of tag-identification system could be used, as well. Of course, voice commands can also be used to facilitate personnel replacements in the medical environment.

Image processing demands, especially for 3-D imaging, can place an enormous burden on the operating computer system 1038. As mentioned hereinabove, the system 1308 can be distributed across two or more computers as a multi-computer system to supply the processing power for 3-D image processing. The disclosed imaging system software can then be distributed across the multi-computer system for the exchange of data needed for ultimately making decisions for human-machine interaction.

The system 1308 can also employ a bi-directional interaction scheme to confirm selection of all gesture, and gesture/voice actions. For example, if the user initiates a user command for starting a session, and the system 1308 responds with a signal that further requires a user response to confirm that a session is to begin. The confirmation respond can be in the form of a duplicate gesture and/or voice command. Obviously, the number and combination of gestures and voice commands that can be employed singly or in combination in accordance with the present system are numerous.

The system 1308 also includes audio input capabilities such that not only voice signals can be received and processed, but clicking sounds, pitch-related sounds, etc., and other distinctive audio signals can be employed to further extend the number of inputs for controlling the system 1308. Such alternative inputs can be input through the portable microphone system 1328 worn by at least one medical person in the operating room. Moreover, additional haptics inputs can be employed by providing a suit or vest with various touch or pressure points to augment the number of signals for controlling the system 1308. Thus, the wrist, forearm, and other appendage points can be used to initiate and send signals from the suit through a wireless remote pressure point transmission system, made part of the wireless voice communication system 1328, for example.

Referring now to FIG. 18, there is illustrated a flowchart of a process from the perspective of the person for using the system of FIG. 17. At 1800, the user performs a calibration process that comprises associating (or mapping) a number of gestures in the form of hand poses and movements, head movements, and/or voice commands with user commands. This also includes using voice commands singly to control the operation computer system or in combination with the gestures to do so. The calibration process can be performed well in advance of use in the operating room, and updated as the user chooses to change movements and/of voice signals with user commands. At 1802, the person initiates a session with the computer system using one or more gestures. At 1804, the person then inputs one or more of the user commands using gestures to control operation of the computing system. At 1806, the person terminates the session using one or more of the gestures. The process then reaches a Stop block.

Referring now to FIG. 19, there is illustrated a flowchart of a process from the perspective of the system of FIG. 17. At 1900, the calibration process occurs for a user where the user presents one or more hands and, hand poses, and orientations to the imaging system for capture and association with a given user command. The system then maps the images to the user command. This occurs for a number of different commands, and completes the calibration phase for that user. At 1902, the user presents one or more gestures that are captured and processed by the system for user commands. At 1904, the system determines if the processed user command(s) indicate that a session is to be started with that user. If NO, flow is back to the input of 1902 to continue to process of

receiving and interpreting gestures and/or voice signals. If YES, flow is to 1906 to identify the user. This can be *via* a triangulation system that determines the location of the source of the gestures. In one implementation, a glove of the medical person includes an RF device, or similar device that is detectable by the system for the purpose of determining the source of the gesture signals. At 1908, the user profile of calibration data is activated for use. Of course, on any given operation, the operating staff are assigned, such that the log-in names of the doctors and assistants can be entered prior to beginning the operation. Thus, the user profiles are already activated for processing.

At 1910, the gestures are imaged, received, and processed to execute the corresponding the user command(s). At 1912, where a classifier is employed, the classifier tracks, processes gesture images, compares the images, and updates user gestures characteristics associated with the particular user command. At 1914, the computer system determines if the session has completed. If NO, flow is back to the input of 1910 to continue to process gestures into user commands. If YES, flow is to 1916 to terminate the user session. The process then reaches a Stop block. Of course, from 1916, flow can be brought back to the input of 1902 to continue to process gestures or to prepare for another session, which could occur many times during the operating room event.

Referring now to FIG. 20, there is illustrated a medical environment 2000 in which a 3-D imaging computer control system 2002 is employed with the remote control device 1404 to process hand (or body) gestures and control the system 1308 in accordance with the present invention. The imaging and image processing capabilities of the 3-D imaging system 1308 and the head-mounted wand 1404 can be employed in combination to further enhance the hands-free capabilities of the present invention. Moreover, the wireless vocalization system 1328 can further be used to augment control of the system 1308. As indicated previously, the wand electronics can be repackaged for use in many different ways. For example, the packaging can be such that the wireless system is worn on the wrist, elbow, leg, or foot. The system 1308 can be used to image both the gestures of the person 1306 and the orientation of the wand 1404 to provide more accurate human-machine interaction and control. Each of the systems have been described herein, the details of which are not repeated here for the purpose of brevity.

Sample gestures, voice commands and gaze signals used in the system 2002 are described hereinbelow.

Referring now to FIG. 21A, there is illustrated sample one-handed and two-handed gestures that can be used to control the operation computing system in accordance with the present invention. At 2100, two closed fists (left and right) can be programmed for imaging and interpretation to cause axis control. At 2102, the right hand in a pointing pose can be used in two orientations, a vertical orientation followed by a sideways clockwise rotation, the combination of which can be programmed for imaging and interpretation to tilt a selected axis a predetermined number of degrees, and keep tilting the axis in stepped increments. At 2104, a continuation of the gestures of 2102 in reverse where, the sideways clockwise rotation is reversed to a counterclockwise rotation followed by the vertical orientation, the combination of which can be programmed for imaging and interpretation stop axis tilting, and maintain at the current tilt angle. At 2106, a right-handed two-fingers-raised pose can be used to rotate an image about an existing axis. Note that the image can be x-rays of the patient, MRI (Magnetic Resonance Imaging) frames, etc. At 2108, the thumb and pointing finger pose of the right hand can be used to rotate an image about a vertex point.

Referring now to FIG. 21B, there is illustrated additional sample one-handed gestures and sequenced one-handed gestures that can be used to control the operation computing system in accordance with the present invention. At 2110, an open right hand with fingers tightly aligned can be used to initiate a zoom-in feature such that the zoom-in operation continues until the gesture changes. At 2112, a right hand where the thumb and pinky finger are extended can be used to initiate a zoom-out feature such that the zoom-out operation continues until the gesture changes. At 2114, a sequence of right-hand gestures are used to select an image for x,y translation, and then to translate the image up and to the right by a predefined distance or percentage of available viewing space on the display. Here, the right hand is used to provide an open hand plus closed fist plus open hand, and then move the open hand up and to the right a short distance. This can be recognized and interpreted to perform the stated function of axis translation in an associated direction. At 2116, a sideways pointing pose plus a counterclockwise motion is programmed for interpretation to rotate the object in the horizontal plane. At 2118, the

same hand pose plus a circular motion in the opposite direction can be programmed to rotate the object in the vertical plane. Note, however, that the hand pose is arbitrary, in that it may be more intuitive to use a hand pose where one or ore of the fingers point upward. Moreover, the gesture, itself is also arbitrary, and is programmable according to the particular desires of the user.

Referring now to FIG. 21C, there is illustrated additional sample one-handed gestures that can be used to control the operation computing system in accordance with the present invention. At 2120, a three-finger open with index and thumb touching of the right hand can be used to impose a triaxial grid on a 3-D image. At 2122, a right-handed single pointing-finger pose can be used to select the x-axis; a right-handed two-finger pose can be used to select the y-axis; and, a right-handed three-finger pose can be used to select the z-axis. At 2124, a pinky-finger pose can be used to stop, start and loop videos on the system 1308. It is also possible using various "structure-from-motion" techniques to track arbitrary points on the hand, and over time, deduce the change in 3-D orientation of the object in such a way that the user need not adopt some predefined pose. In this case, however, the user can enter the 3-D rotation mode by another method.

At 2126, rotation of the pinky-finger pose in a clockwise direction while facing the system 1308 can be used to control intensity of the monitor, and volume on/off control and amplitude. These are being grouped of brevity, since, for example, the pinky-finger pose and/or rotation can be mapped to any one of the functions described. At 2128, an open hand gesture in a clockwise rotation can be used to rotate an image about an axis according to the speed of movement of the open hand, such that when the hand stops, the axis rotation also stops, and starts when hand movement starts.

Referring now to FIG. 21D, there is illustrated additional sample one-handed gestures used in combination with voice commands that can be used to control the operation computing system in accordance with the present invention. At 2130, the open hand pose plus a voiced "ZOOM" command can be used to zoom in on a displayed image until the gesture changes or a different command is voiced. At 2132, the thumb and pinky finger extended pose plus a voiced "ZOOM" command can be used to zoom out on a displayed image until the gesture changes, or a different command is voiced. Depth information can also be used, e.g., moving closer would trigger a zoom-in function.

Alternatively, when zoom is invoked, movement in depth can control the zoom value continuously.

At 2134, a left-handed open hand pose in a sideways orientation plus a voiced "MOVE" command can be used to move a selected image to the right until the gesture changes and stops movement. At 2136, a right-handed open hand pose in a sideways orientation plus a voiced "MOVE" command can be used to move a selected image to the left until the gesture changes and stops movement. At 2138, a closed fist in a circular motion in combination with a "LOUD" voice command can be used to turn audio volume on/off, and control the amplitude during the procedure to listen to the patient's EKG, for example.

Referring now to FIG. 21E, there is illustrated additional sample one-handed gestures used in combination with voice commands and gaze signals that can be used to control the operation computing system in accordance with the present invention. At 2140, the right-hand open-hand pose in combination with a voiced "ZOOM" command while gazing at an image on a first display of the operation computer system will invoke a zoom-in process on the image of the first display until the gesture is changed. At 2142, the thumb and pinky finger extended of the pose of the right hand is used in combination with a voiced "ZOOM" command while gazing in the direction of an image presented on a second display to control the computer system to zoom out on the image of the second display until the gesture changes. At 2144, a left-handed open hand pose in a sideways orientation in combination with a voiced "MOVE" command while gazing at an image on a first display of the operation computer system will invoke a rightward move operation on the image of the first display until the gesture is changed. At 2146, a right-handed open hand pose in a sideways orientation in combination with a voiced "MOVE" command while gazing at an image on a second display of the operation computer system will invoke a leftward move operation on the image of the second display until the gesture is changed. At 2148, a closed right fist in a circular clockwise motion in combination with a voiced "LOUD" command and a gaze in the direction of a graphical interface of an audio control device on a third display of the computer control system results in volume on/off control and amplitude control.

It is to be appreciated that numerous other combinations of hand poses, body gestures, voice commands and gaze orientations can be employed to effect control of the medical operation environment. Only a few samples of the individual and combinatory capabilities are provided herein.

The complementary nature of speech and gesture is well established. It has been shown that when naturally gesturing during speech, people will convey different sorts of information than is conveyed by the speech. In more designed settings such as interactive systems, it may also be easier for the user to convey some information with either speech or gesture or a combination of both. For example, suppose the user has selected an object as described previously and that this object is a stereo amplifier controlled via a network connection by the host computer. Existing speech recognition systems would allow a user to control the volume by, for example, saying "up volume" a number of times until the desired volume is reached. However, while such a procedure is possible, it is likely to be more efficient and precise for the user to turn a volume knob on the amplifier. This is where the previously described gesture recognition system can come into play. Rather than having to turn a physical knob on the amplifier, the user would employ the pointer to control the volume by, for example, pointing at the stereo and rolling the pointer clockwise or counterclockwise to respectively turn the volume up or down. The latter procedure can provide the efficiency and accuracy of a physical volume knob, while at the same time providing the convenience of being able to control the volume remotely as in the case of the voice recognition control scheme. This is just one example of a situation where gesturing control is the best choice, there are others. In addition, there are many situations where using voice control would be the best choice. Still further, there are situations where a combination of speech and gesture control would be the most efficient and convenient method. Thus, a combined system that incorporates the previously described gesturing control system and a conventional speech control system would have distinct advantages over either system alone.

To this end, as indicated hereinabove, the present invention includes the integration of a conventional speech control system into the gesture control and pointer systems which results in a simple framework for combining the outputs of various modalities such as pointing to target objects and pushing the button on the pointer,

pointer gestures, and speech, to arrive at a unified interpretation that instructs a combined environmental control system on an appropriate course of action. This framework decomposes the desired action (*e.g.*, "turn up the volume on the amplifier") into a command (*i.e.*, "turn up the volume") and a referent (*i.e.*, "the amplifier") pair. The referent can be identified using the pointer to select an object in the environment as described previously or using a conventional speech recognition scheme, or both. The command may be specified by pressing the button on the pointer, or by a pointer gesture, or by a speech recognition event, or any combination thereof. Interfaces that allow multiple modes of input are called multimodal interfaces. With this multimodal command/referent representation, it is possible to effect the same action in multiple ways. For example, all the following pointing, speech and gesture actions on the part of the user can be employed in the present control system to turn on a light that is under the control of the host computer:

1. Say "turn on the desk lamp";
2. Point at the lamp with the pointer and say "turn on";
3. Point at the lamp with the pointer and perform a "turn on" gesture using the pointer;
4. Say "desk lamp" and perform the "turn on" gesture with the pointer;
5. Say "lamp", point toward the desk lamp with the pointer rather than other lamps in the environment such as a floor lamp, and perform the "turn on" gesture with the pointer; and
6. Point at the lamp with the pointer and press the pointer's button (assuming the default behavior when the lamp is off and the button is clicked, is to turn the lamp on).

By unifying the results of pointing, gesture recognition and speech recognition, the overall system is made more robust. For example, a spurious speech recognition event of "volume up" while pointing at the light is ignored, rather than resulting in the volume of an amplifier being increased, as would happen if a speech control scheme were being used alone. Also, consider the example given above where the user says "lamp"

while pointing toward the desk lamp with the pointer rather than other lamps in the environment, and performing the "turn on" gesture with the pointer. In that example, just saying lamp is ambiguous, but pointing at the desired lamp clears up the uncertainty. Thus, by including the strong contextualization provided by the pointer, the speech recognition may be made more robust.

The speech recognition system can employ a very simple command and control (CFG) style grammar, with preset utterances for the various electronic components and simple command phrases that apply to the components. The user wears a wireless lapel microphone to relay voice commands to a receiver which is connected to the host computer and which relays the received speech commands to the speech recognition system running on the host computer.

While various computational frameworks could be employed, the multimodal integration process employed in the present control system uses a dynamic Bayes network that encodes the various ways that sensor outputs may be combined to identify the intended referent and command, and initiate the proper action.

The identity of the referent, the desired command and the appropriate action are all determined by combining the outputs of the speech recognition system, gesture recognition system and pointing analysis processes using a dynamic Bayes network architecture. Bayes networks have a number of advantages that make them appropriate to this task. First, it is easy to break apart and treat separately dependencies that otherwise would be embedded in a very large table over all the variables of interest. Secondly, Bayes networks are adept at handling probabilistic (noisy) inputs. Further, the network represents ambiguity and incomplete information that may be used appropriately by the system. In essence, the Bayes network preserves ambiguities from one time step to the next while waiting for enough information to become available to make a decision as to what referent, command or action is intended. It is even possible for the network to act proactively when not enough information is available to make a decision. For example, if the user doesn't point at the lamp, the system might ask which lamp is meant after the utterance "lamp".

However, the Bayes network architecture is chosen primarily to exploit the redundancy of the user's interaction to increase confidence that the proper action is being

implemented. The user may specify commands in a variety of ways, even though the designer specified only objects to be pointed to, utterances to recognize and gestures to recognize (as well as how referents and commands combine to result in action). For example, it is natural for a person to employ deictic (pointing) gestures in conjunction with speech to relay information where the speech is consistent with and reinforces the meaning of the gesture. Thus, the user will often naturally indicate the referent and command applicable to a desired resulting action via both speech and gesturing. This includes most frequently pointing at an object the user wants to affect.

The Bayes network architecture also allows the state of various devices to be incorporated to make the interpretation more robust. For example, if the light is already on, the system may be less disposed to interpret a gesture or utterance as a "turn on" gesture or utterance. In terms of the network, the associated probability distribution over the nodes representing the light and its parents, the Action and Referent nodes, are configured so that the only admissible action when the light is on is to turn it off, and likewise when it is off the only action available is to turn it on.

Still further, the "dynamic" nature of the dynamic Bayes network can be exploited advantageously. The network is dynamic because it has a mechanism by which it maintains a short-term memory of certain values in its network. It is natural that the referent will not be determined at the exact moment in time as the command. In other words a user will not typically specify the referent by whatever mode (*e.g.*, pointing and/or speech) at the same time he or she relays the desired command using one of the various methods available (*e.g.*, pointer button push, pointer gesture and/or speech). If the referent is identified only to be forgotten in the next instant of time, the association with a command that comes after it will be lost. The dynamic Bayes network models the likelihood of a referent or a command applying to future time steps as a dynamic process. Specifically, this is done *via* a temporal integration process in which probabilities assigned to referents and commands in the last time step are brought forward to the current time step and are input along with new speech, pointing and gesture inputs to influence the probability distribution computed for the referents and commands in the current time step. In this way, the network tends to hold a memory of a command and referent that decays over time, and it is thus unnecessary to specify the command and

referent at exactly the same moment in time. In one example, this propagation occurred four times a second.

Note that although a previous description was centered on an operating room environment, the present invention has application in many other environments where data access and presentation is beneficial or even necessary to facilitate a man-machine interface.

What has been described above includes examples of the present invention. It is, of course, not possible to describe every conceivable combination of components or methodologies for purposes of describing the present invention, but one of ordinary skill in the art may recognize that many further combinations and permutations of the present invention are possible. Accordingly, the present invention is intended to embrace all such alterations, modifications, and variations that fall within the spirit and scope of the appended claims. Furthermore, to the extent that the term "includes" is used in either the detailed description or the claims, such term is intended to be inclusive in a manner similar to the term "comprising" as "comprising" is interpreted when employed as a transitional word in a claim.